COMPUTATIONAL MODELING AND ANALYSIS OF SINGLE-CELL SIGNALING DYNAMICS IN HETEROGENEOUS CELL POPULATIONS

A Dissertation
Presented to
The Academic Faculty

By

James D. Wade

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Wallace H. Coulter Department of Biomedical Engineering

Georgia Institute of Technology
Emory University

May 2019

COMPUTATIONAL MODELING AND ANALYSIS OF SINGLE-CELL SIGNALING DYNAMICS IN HETEROGENEOUS CELL POPULATIONS

Approved by:

Dr. Eberhard O. Voit, Advisor
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology & Emory University School of Medicine

Dr. Bernd Bodenmiller
Department of Quantitative Biomedicine
University of Zurich

Dr. Barbara D. Boyan
College of Engineering
Virginia Commonwealth University

Dr. Melissa L. Kemp
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology & Emory University School of Medicine

Dr. John F. McDonald
School of Biology
Georgia Institute of Technology

Dr. Peng Qiu
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology & Emory University School of Medicine

Date Approved: February 1, 2019

To Dr. Andrew J. "Andy" Kozar

Who inspired me first as an athlete

then, as a scholar

and always, as a human being.

ACKNOWLEDGMENTS

Many people deserve acknowledgement for their support during my journey to finding, beginning, continuing and finishing this dissertation. I am genuinely grateful for those listed below, and very many others who are not.

I so owe much to my doctoral advisor Dr. Eberhard O. Voit, whose mentorship has gone above and beyond academic pursuits, and whose support, patience and trust have enabled me to pursue this work. To my original co-advisor Dr. Barabara D. Boyan, who first grabbed me while on an exploratory visit, supported my transition out of athletics, introduced me to biology and encouraged my continued explorations. To my de-facto co-advisor Bernd Bodenmiller, who welcomed me to his young lab, has supported my journey with enthusiasm and has always provided a space that encourages shooting for the moon. To the rest of my committee members, Dr. Melissa L. Kemp, Dr. John McDonald and especially Dr. Peng Qiu, who have all provided thoughtful input and guidance throughout this work.

I have had many colleagues and friends across the Voit, Bodenmiller and Boyan labs, as well as across Georgia Tech, Emory, University of Zurich and ETHZ whose thoughtful discussion has been invaluable to this work. Drs. Xiaokang Lun and Nevena Zivanovic, who taught me how to work at the bench, spent long nights with me in lab, made major contributions to this work and have become dear friends. Also, Dr. Luis Fonseca and Vito Zanotelli, friends and colleagues who both have provided essential discussion and feedback on modeling and biology. Andrea Jacobs, who was patient and understanding in my early days at the bench, and who was always helpful. It would take many pages to describe the full appreciation I have for these and so many others during my doctoral studies, both as intellectual contributors and as friends.

This dissertation has spanned multiple life transitions: from athlete to academic; from single to married; from large child to father. Many, many people have played roles in these

transitions, not least those above, but there are a few more I would like to mention by name. Dr. Andrew J. Kozar, who planted the seed of possibility while I was but a child. Rafal Smolen, my coach for so many years, for the role he has played in me becoming myself. Dr. Ozlem Ergun, my bachelors research advisor, who inspired, guided and encouraged me both as a researcher and a person. She still inspires me. Dr. Larry Jacobs and Kay Kinard, who helped me search across fields, which led to this work and becoming a biomedical engineer. Dr. Karen Adams, who contributed so much to the fellowship applications that enabled this work and, by getting me to Switzerland, to the start of my family.

Finally, but not lastly, is family. To my parents: there is too much to say. Thank you both. For all the years. To Marianne and Werner, thank you for the support in the final days. To Pascal, because, you know. To my wife and son, Laura and Lucien: I love you both.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS OR ABBREVIATIONS

AKT	protein kinase B.
BP-R^2	binned pseudo R-squared.
CMV	cytomegalovirus.
C/N	cytoplasmic-to-nuclear ratio.
CSC	cancer stem cell.
EGF	epidermal growth factor.
EGFR	epidermal growth factor receptor.
EMD	earth movers distance.
EMT	epithelial mesenchymal transition.
ERK	extracellular signal-regulated kinase.
FRET	Förster resonance energy transfer.
HEK	human embryonic kidney.
JNK	c-Jun N-terminal kinase.
K-S	Kolmogorov-Smimov.
KTR	kinase translocation reporter.
MAPK	mitogen-activated protein kinase.
MCB	mass-tag cellular barcoding.
MERFISH	multiplexed error-robust fluorescence in situ hybridization.
MLE	maximum likelihood estimation.
MMD	maximum mean discrepancy.
MST	minimal spanning tree.
ODE	ordinary differential equation.
PI3K	phosphatidylinositol-4,5-bisphosphate 3-kinase.
PLA	proximity ligation assay.
RKHS	reproducing kernel Hilbert space.
RSK	p90 ribosomal s6 kinase.
RTK	receptor tyrosine kinase.

S6	ribosomal protein S6.
SPKF	sigma-point Kalman filter.
TOF-MS	time-of-flight mass spectrometry.

SUMMARY

Cell signaling pathways are complex biochemical systems at the core of cellular information processing. The dynamics of these signaling systems in response to internal and extracellular cues plays a critical role for proper cell functioning. While we have learned much about signaling at the cell population level, no two cells are the same, and cell-to-cell variability can have complex and important consequences for signaling in both individual cells and the cell population as a whole. In many contexts, cells perform essentially identical functions despite their differences, whereas in other contexts, especially in cancer, cell-cell differences in state propagate to differences in function.

The overall goal of this dissertation was the creation of mathematical and computational tools for the study of cell-to-cell variation in signaling and to use these tools to increase our understanding of when single cell differences do, or do not, make a meaningful difference. To address this goal we designed new methods of single-cell analysis, including a computational framework termed single-cell ordinary differential equation modeling (SCODEM) that overcomes the prior experimental trade-off between continuous and multiplexed single-cell measurements of signaling. We tested SCODEM against increasingly demanding datasets, which were all represented in a satisfactory fashion. After the initial analysis of cell-to-cell variability, we analyzed targeted inhibition, protein overexpression and an epithelial-mesenchymal transition. Throughout this process, we provided illustrative examples of how our modeling framework may be used to identify operating principles and limits of signaling systems, which is a first step toward proposing novel therapeutic targets.

The work presented here provides new tools for analyzing cellular heterogeneity and increases our understanding of how differences in cell state effect function by showing intracellular signaling is primarily deterministic at the single cell level. The application of these tools to the dramatic phenotype shift during an epithelial-mesenchymal transition in murine breast cancer cells confirmed that stochasticity plays a much smaller role than had been assumed and that cells modulate signaling without the need of rewiring their signaling network.

CHAPTER I

INTRODUCTION

1.1 Cell-to-cell Variation

Cells may be considered as biochemical systems whose states can be broadly defined by the relative locations and concentrations or abundances of molecular components at a point in time. Single-cell measurement technologies have shown that individual cells, even within what may classically be considered a homogeneous cell population or sample, exhibit considerable variation in state (Farlik et al., 2015; Macosko et al., 2015; Sachs et al., 2005; Wang and Bodovitz, 2010). This variation among cells within a population has meaningful consequences for the functional responses of individual cells. For example, cell-to-cell variation in state leads divergent receptor-induced apoptotic responses in cells from genetically identical populations (Spencer et al., 2009). In cancer and other diseases, genetic alterations modify the expression or function of cellular components, which can result in increased cellular variation and create cells with aberrant functional responses to microenvironmental inputs compared to healthy cells (Cohen et al., 2008; Altschuler and Wu, 2010). Microenvironmental changes associated with disease can further increase heterogeneity in cell states and inputs and, consequently drug responses (Meacham and Morrison, 2013; Burrell and Swanton, 2014). However, variation in cell state alone is not a sufficient prerequisite for differential responses (Bendall et al., 2011), and despite the clear role of cell-to-cell variation in the generation of differential cell fates, a precise definition of when, where and what variation plays a functional role remains an open question across many contexts (Altschuler and Wu, 2010; Snijder and Pelkmans, 2011).

1.2 Intracellular Signaling

This dissertation focuses on the role of cellular variation in the specific subsystem of intracellular signaling. Intracellular signaling pathways are complex biochemical systems at the

core of cellular information processing, and the dynamics of these signaling systems in response to extracellular cues plays a critical role for proper cell functioning (Dolmetsch et al., 1997; Kholodenko, 2006; Selimkhanov et al., 2014). Signals are transduced by modulating the enzymatic activities and local concentrations of signaling mediators such as protein kinases. Variation in the relative expression of signaling components can lead to differential signaling responses and cell fates (Cohen-Saidon et al., 2009). As a consequence that is of particular importance here, many cancers are caused by aberrant expression or function of signaling proteins (Sever and Brugge, 2015) that alter normal input-output signaling responses, and the increased heterogeneity in cell states within tumors can result in drug resistance (Dagogo-Jack and Shaw, 2018; Shaffer et al., 2017).

1.3 Methods to Study Cell-to-Cell Variation in Signaling

While cell-to-cell variation plays a clear role in cell responses to signaling inputs (Shaffer et al., 2017; Spencer et al., 2009), the prediction and analysis of single-cell signaling responses remains a significant challenge. This is due, at least in part, to the complex interconnected structure of signaling networks and the generally non-linear nature of the associated reaction dynamics (Papin et al., 2005; Hengl et al., 2007). Reverse engineering such complex systems requires simultaneous observations of multiple components, for instance, in the form of multivariate or 'multiplexed' measurements over time (Kolitz and Lauffenburger, 2012; Spiller et al., 2010). However, all current experimental methods face a trade-off between the ability to observe the state of many system variables as a snapshot at an individual point in time, which alone cannot characterize response, and the option of monitoring the state of one or few variables continuously in time, which cannot characterize how system components interact to drive response.

Mathematical and computational modeling approaches can complement experimental approaches by providing a formal framework for the representation and analysis of signaling systems and are in principle able to bridge the gap between snapshot data and system responses (Spiller et al., 2010). However, they require quite detailed *a priori* knowledge of the involved signaling networks, as well as very comprehensive data. As this dissertation

demonstrates, these data must be obtained at a single-cell level to reveal the information necessary to characterize intercellular variability. The following subsections provide a brief review of experimental and computational methods used to characterize single-cell variation in intracellular signaling.

1.3.1 Experimental Methods

Single-cell measurements are necessary to characterize cell-to-cell variation in a sample and can generally be characterized as continuous, i.e., longitudinal measurements of live cells, or as snapshots, which are fixative or destructive in nature and observe a cell at only an individual point in time.

Continuous Methods

Live-cell methods typically use dyes, for example, to monitor calcium signaling (Chung et al., 2011), or genetically encoded sensors for protein modifications to study signaling. Förster resonance energy transfer (FRET) is a common genetically encoded sensor that relies on proximity-based energy transfer between two fluorophores and has been used to image the activation and localization of signaling proteins at the single-cell level (Grant et al., 2008; Ryu et al., 2015). Challenges in applying FRET to study signaling include low signal-to-noise ratio and inaccurate characterization of signal degradation due to stable 'on' configurations (Komatsu et al., 2011; Regot et al., 2014). Additionally, FRET methods require the use of two separate measurement channels (wavelengths), which reduces the ability to multiplex measurements as most microscopes are limited to four channels. Another type of genetically encoded fluorescent biosensor is the Kinase Translocation Reporter (KTR). KTRs are constructed using nuclear targets of kinases and, upon phosphorylation by their target kinase, translocate from the nucleus to the cytoplasm. KTRs have become popular due to the straightforward calculation of signal, cytoplasmic to nuclear ratio (C/N), and their use of only a single measurement channel, which enables multiplexing up to three such reporters on conventional microscopes (Regot et al., 2014). A major drawback is that KTRs cannot be used to measure the activities of kinases that do not translocate to the nucleus upon activation. Compared to FRET, KTRs have been shown to be less sensitive at

low signaling inputs with slightly delayed dynamics of activation, but improved inactivation dynamics and larger dynamic range (Gillies et al., 2017). Finally and most importantly, both FRET and KTR sensors can suffer from a limited dynamic range and, due to the necessity of genetic expression, have limited direct applicability to characterizing signaling responses in primary samples.

Snapshot methods

Compared to live-cell methods, experimental single-cell snapshot methods are better able to quantify the systems-level state of intracellular signaling by measuring more components simultaneously. Generally, these methods rely on antibodies to measure the relative abundances and post-translational states of signaling proteins. While microfluidic techniques have been used to develop single-cell western blots, multiplexing is relatively low at 11 channels and the detection limit is relatively high at 30,000 molecules (Hughes et al., 2014). The most common multiplexed single-cell snapshot methods applied to signaling studies are flow cytometry, and more recently, mass cytometry (Sachs et al., 2005; Bandura et al., 2009; Bodenmiller et al., 2012). Flow cytometry measures antibodies labeled with fluorophores and has been used to measure up to 17 molecular markers simultaneously (Perfetto et al., 2004). Due to technical challenges such as spectral spillover between measurement channels, however, flow cytometry is usually limited in dimensionality to around 10 markers. Mass cytometry, by contrast, uses antibodies labeled with isotopes of heavy metals not otherwise present in biological samples. Individual cells are atomized, ionized and the isotopes are separated and measured using time-of-flight mass spectrometry (TOF-MS). The high specificity of TOF-MS results in greatly reduced spillover between channels and routinely enables the simultaneous measurement of up to 50 markers (Bendall et al., 2012; Spitzer and Nolan, 2016). Currently, mass cytometry represents the state of the art for experimentally characterizing intracellular signaling at a systems level; details are shown in Chapter 2 (Lun et al., 2017) of this dissertation.

1.3.2 Computational Methods

Computational methods have been used for assessing bulk population-level data, in an attempt to overcome the trade-off between continuous and snapshot measurements, and to analyze the complex non-linear nature of intracellular signaling systems. In the case of snapshot measurements, a mathematical model of the underlying system is reverse-engineered and used to infer the characteristics of continuous trajectories. The underlying biochemical reactions of cell signaling networks are most commonly modeled by a system of deterministic ordinary differential equations (ODEs). Although biochemical reactions are fundamentally stochastic in nature, the use of deterministic models in signaling is generally justified by the presence of kinases and other reaction components in abundances (Geiger et al., 2012) sufficient to smooth out the stochastic variability in reactions (Goutsias, 2007; Filippi et al., 2016). Thus, forward-simulation of the ODE model generates continuous trajectories of state variables, such as the phosphorylation states of signaling proteins on a population level. Snapshots of simulated trajectories are then compared to experimental measurements to determine model quality and infer model parameters.

Computational Methods of Modeling Heterogeneity

Population-level models have been quite successful in the past but, by their nature, cannot take cell-to-cell variations in a sample into account. This failure to quantify variation in signaling components is intrinsic and mandates the development of entirely different approaches based on single-cell measurements (Bronstein et al., 2015), and especially snapshot measurements that permit highly multiplexed observations. Current methods for assessing variation in signaling dynamics use parametric distributions to represent single-cell snapshot data (Hasenauer et al., 2011, 2014; Filippi et al., 2016; Loos et al., 2018). Model simulations are used to determine trajectories of both the population mean and variance, which describe the most basic features of the distribution shape and statistically quantify the appropriateness of the model quality. The most recent methods (Filippi et al., 2016; Loos et al., 2018) use a sigma-point, or sigma-point Kalman filter (SPKF), approach (Van der Merwe, 2004) to approximate the time-dependent change of distribution parameters.

Specifically, given a model $\mathbf{g}$, the change in distribution parameters is estimated by the optimal choice of a set of sigma-points, which are points selected from the distribution and weighted to characterize its mean and covariance (see Figure 1.1).

The use of parametric distributions offers increased information but complicates model inferences by adding parameters that must be inferred from data. For example, if a sample contains more than one meaningful population, the modeler must first determine the number of (sub)populations and their associated mixing and shape parameters before pursuing the original task of inferring a reaction model to describe population dynamics. These and other aspects of cell-cell heterogeneity have the consequence that parametric descriptions rapidly become computationally and statistically intractable. In the case of cancer, which is of primary interest here, typical samples contain complex mixtures of subpopulations (Chevrier et al., 2017), thus requiring complex mixture distributions. The number of mixture distribution parameters grows as a polynomial function of model state variables, that is, the number of signaling network components, and quickly leads to an infeasible inference problem. Furthermore, distributional parametric methods do not explicitly consider single-cell trajectories. Due to these and other limitations, and despite the gradually increasing ability of producing or acquiring multiplexed single-cell datasets, parametric methods of modeling heterogeneity have only been applied to experimental data corresponding to states of one or two dimensions (Filippi et al., 2016; Loos et al., 2018). This restriction is unacceptable for our purposes, which suggests abandoning parametric for non-parametric methods that we discuss below.

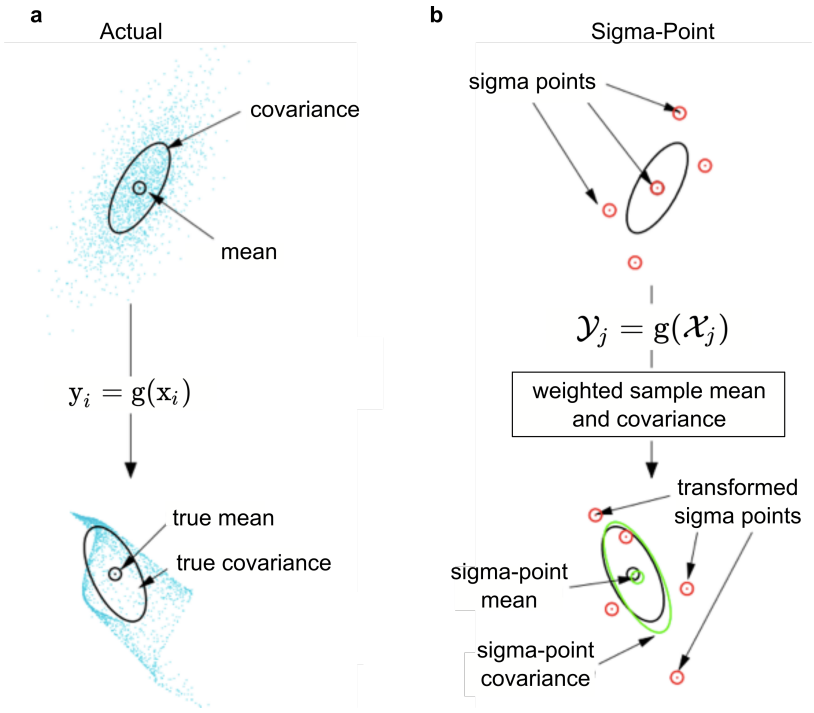


Figure 1.1: Illustration of sigma-point approximation for mean and covariance compared to sampling. Propagation of a random variable x by a nonlinear function g such that y = g(x). **(a)** Direct propagation of many samples x_i, given by $y_i = g(x_i)$, illustrates the true distribution. **(b)** Propagation of sigma-points χ, given by $\mathcal{Y})i = g(\mathcal{X}_j$, that are used to estimate mean and covariance. Figure adapted from (Van der Merwe, 2004)

1.4 Comparing Multivariate Distributions

Model fitting generally amounts to minimizing the difference between experimental observations and results generated by model simulations. In the case of a bulk-population measurement at a time point, for example, the goal is often to minimize the L^2 norm, that is, the squared Euclidean distance, between each experimental observation and the corresponding model simulation result at the given time point. In the case of multiplexed single-cell data at a time point, this goal may be extended to minimizing the difference between the measured and the simulated multivariate distributions. Notably, this extension implies the assumption that accounting for additional distribution features beyond the mean will provide additional information on the structure and function of the system of interest. Current methods that approximate observations by parametric distributions have use maximum likelihood estimation (MLE) to identify such model parameters that model

simulations generate a set of distribution parameters that maximize the likelihood of concurring with the data for each time point (Filippi et al., 2016; Loos et al., 2018). Specifically, these methods have considered the mean and variance (Filippi et al., 2016) or, very recently, mean and covariance (Loos et al., 2018) parameters of the distributions. Advantageously, the gradient of the likelihood function for normal and log-normal distributions can be solved analytically, which speeds up many optimization methods, as demonstrated in Loos et al. (2018). Parametric methods, however, become very cumbersome when the chosen distributions are no longer normal or log-normal. Furthermore, the determination of the precise number, mixing and form of distributions remains a challenging problem, because these distributions are more often than not unknown in biological samples.

In contrast to parametric statistics, nonparametric 'distribution-free' statistical tests are constructed without assumptions regarding the parametric form of the distributions considered. Thus, the difference between two samples consisting of complex mixtures of subpopulations may be characterized without the necessity to define the mixtures *a priori*. In this work, we present a reaction modeling framework based on simulating trajectories for many individual cells that, together, represent empirical multivariate distributions analogous to experimental measurements at corresponding time points (see Chapter 3). In this context, each comparison between simulated and measured multivariate distributions can be reduced to a two-sample test, using distribution-free statistics. Numerous distribution-free two-sample tests for multivariate distributions have been proposed over the past decades. A non-exhaustive list may begin with an extension of quantile-quantile plots to a multivariate quantile-quantile test (Dhar et al., 2014). Non-bipartite matching uses the minimal distance between pairs of points across distributions (Rosenbaum, 2005). Multivariate generalizations of the Smirnov maximum deviations test and the Wald-Wolfowitz runs test are based on using the minimal spanning tree (MST) of sample points as a generalization of the univariate sorted list (Friedman and Rafsky, 1979). An approximate extension of the Kolmogorov-Smirnov (K-S) test to the multivariate case was proposed in Justel et al. (1997). The multivariate earth movers distance (EMD) method first finds an efficient compartmentalization, which corresponds to binning in a histogram, of the probability mass

distribution and, subsequently, computes the minimal cost to transform one distribution into the other by moving probability mass, based on a distance metric such as the L^2 norm (Rubner et al., 2000). The final example uses the maximum mean discrepancy (MMD), which is the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space (RKHS) (Gretton et al., 2012a). Notably, both the approximate K-S test and EMD are special cases of the MMD taken over function classes other than the RKHS (Gretton et al., 2012a); all of these tests are fundamentally related to the idea of the K-S test, which has been in use for many decades.

For the purpose of fitting reaction model dynamics to time course data, an appropriate test must be both computationally efficient and discriminatory. Efficiency is necessary as model fitting generally requires very large numbers of comparisons. Discrimination must be considered for each test as an associated functional definition of distance or similarity between distributions, which may perform better or worse in guiding a parameter search algorithm. With these considerations in mind, MMD has been shown to perform well compared to other methods across a broad range of distribution shapes and incurs relatively low computational cost at the order of $O\big((n+m)^2\big)$, given sample numbers n and m from each distribution; indeed, it may even be approximated in linear time, if necessary. Thus, MMD becomes the basis for the non-parametric single-cell reaction modeling framework we develop in Chapter 3.

1.5 Epidermal Growth Factor Signaling

In this work, we study signaling in the epidermal growth factor receptor (EGFR) pathway. EGFR signaling controls cell growth, motility, survival, differentiation, and metabolism (Citri and Yarden, 2006). The EGFR signaling network is highly studied and the reaction structure is described in the literature in multiple levels of detail (see, for example, Kholodenko (2006)). EGFR signaling network proteins (e.g., EGFR, HER2, ERK, and AKT) are affected by gene copy number alterations that deregulate protein abundances and/or functions in a number of cancer types (Govindarajan et al., 2007; Eralp et al., 2008a; Han et al., 2015). Intriguingly, targeted treatments for EGFR network-related cancers have had more

limited success than expected, which is due, at least in part, to cell-cell variation in tumor cell states (Wellbrock and Arozarena, 2016; Caunt et al., 2015). The vast body of prior knowledge and the functional significance of EGFR signaling provide a solid basis upon which to test our methodological developments. We take a broad approach to studying how variation in cell states affects signaling in the EGFR network by considering populations of wild-type cells, cell populations that include a broad and continuous range of signaling protein overexpression and, finally, cell populations both before and after a phenotypic transition (see next section). Relevant descriptions of the EGFR signaling network itself may be found in the introductions of subsequent chapters.

1.6 Epithelial Mesenchymal Transition

The epithelial-to-mesenchymal transition (EMT) is a developmental program that naturally occurs in embryogenesis and wound healing. During this transition, polarized epithelial cells de-differentiate into a mesenchymal phenotype characterized by loss of cell-cell adhesion junctions, increased capacity for migration and invasion, and resistance to apoptosis (Fu et al., 2018). In cancer, EMT plays a major role in the generation of small cancer stem cell (CSC) subpopulations that exhibit both increased drug resistance and the ability to regenerate tumors post treatment (Shibue and Weinberg, 2017; Du and Shim, 2016). Cells with an EMT signature, for example, have specific resistance to drugs targeting EGFR signaling pathways, such as EGFR and PI3K/AKT inhibitors (Byers et al., 2013).

1.7 Objectives and Summary

The overall goal of this dissertation is to develop mathematical and computational methods for the study of cell-to-cell variation in signaling, and to use these tools to increase our understanding of when single cell differences do, or do not, make a meaningful difference. To address this goal, we take a progressive approach. The complexity of modeling frameworks moves from descriptive statistical measures to mechanistic reaction models, and the range of cellular variation increases from wild-type cells to mixtures of wild-type and protein-overexpressing cells and ultimately to cells before and after the dramatic phenotypic change induced by EMT.

Although cell-to-cell variation is an undisputed observation in signaling systems, our overarching hypothesis is that most of this variation is not random. Expressed differently, our hypothesis is that effective signaling is so important for the cell that the signaling program is primarily deterministic at the single-cell-level, and thus reliably repeatable. Pursuing this hypothesis, we design new methods of single-cell analysis and apply them to increasingly demanding scenarios, in which the apparent heterogeneity increases. Ultimately we show that a minimal mechanistic model with constant structure and nearly identical parameter values is able to explain the dramatic changes in signaling function across different scenarios, including EMT. The model results even pinpoint specific molecular processes that are responsible for the remaining variation, which is incomparably smaller than had been postulated based on earlier data analyses. The fact that a single model can explain substantial changes in signaling function suggests a much higher role of deterministic functioning and a much more reduced role of stochasticity than previously assumed.

The chapters of this dissertation are arranged as follows:

Chapter 2: This chapter describes a novel experimental and statistical framework for high-throughput systems-level analysis of protein expression-dependent effects on signaling. We use this framework and overexpression of 20 protein kinases to study EGFR network signaling in 360 conditions with a panel of 35 antibodies. We identify previously unreported signaling relationships and illustrate effects of altered protein abundance or concentration on signaling dynamics.

Chapter 3: This chapter describes a novel ODE-based computational modeling approach to infer multivariate single-cell signaling dynamics from multiplexed single-cell snapshot data. This method overcomes the experimental trade-off between multiplexing and time-series measurements of cell response. We use our method to study EGF signaling in the MAPK/ERK pathway in the context of both wild-type populations as well as complex mixture populations, generated using the experimental techniques established in Chapter 2. We show how protein overexpression can reveal complex kinetic effects that are sufficient to

explain the altered observed dynamics discussed in Chapter 2. We also show that the variation in signaling responses across wild-type cells is generally much less than the variation in cell states, yet sufficient to drive subpopulation-dependent drug responses. Finally, we provide an illustrative example of how single-cell reaction models may be used to identify novel treatment strategies.

Chapter 4: In this chapter, we use established data-driven network inference methods and the single-cell ODE modeling method developed in Chapter 3 to study EGF signaling in the ERK and AKT pathways before and after EMT. We show that, given observations of appropriate initial cell states, a single reaction-based model of signaling with constant structure and near-constant parameters is sufficient to represent differences in EGF signaling across EMT.

Chapter 5: This final chapter summarizes our contributions to the body of scientific knowledge and discusses future directions.

CHAPTER II

INFLUENCE OF NODE ABUNDANCE ON SIGNALING NETWORK STATE AND DYNAMICS ANALYZED BY MASS CYTOMETRY[1]

2.1 Abstract

Signaling networks are key regulators of cellular function. Although the concentrations of signaling proteins are perturbed in disease states, such as cancer, and are modulated by drug therapies, our understanding of how such changes shape the properties of signaling networks is limited. Here we couple mass cytometry-based single-cell analysis with overexpression of tagged signaling proteins to study the dependence of signaling relationships and dynamics on protein node abundance. Focusing on the epidermal growth factor receptor (EGFR) signaling network in HEK293T cells, we analyze 20 signaling proteins during a one hour EGF stimulation time course using a panel of 35 antibodies. Data analysis with BP-R^2, a measure that quantifies complex signaling relationships, reveals abundance-dependent network states and identifies novel signaling relationships. Further, we show that upstream signaling proteins have abundance-dependent effects on downstream signaling dynamics. Our approach elucidates the influence of node abundance on signal transduction networks and will further our understanding of signaling in health and disease.

2.2 Introduction

Signaling networks are at the core of cellular information processing and transform external signals into cellular responses. Signals are transduced by modulating enzymatic activities mainly via protein phosphorylation, and cells implement sophisticated mechanisms, such as feedback loops, pathway crosstalk, and differential enzyme localization, to integrate signals

[1]This chapter is adapted from: Lun, XK*, Zanotelli, VRT*, Wade, JD*, Schapiro, D, Tognetti, M, Dobberstein, N and Bodenmiller, B. (2017) Influence of node abundance on signaling network state and dynamics analyzed by mass cytometry. ***Nature Biotechnology***, 35(2):164-172. *Conceptualization of analysis, performance of analysis.

and drive cellular processes and physiological outputs. The abundance of individual signaling pathway components (nodes) is central to the activity and output of a signaling network (Wolf-Yadlin, 2006). Changes in node abundance are tightly regulated and control biological programs such as stem cell differentiation and embryogenesis (De Los Angeles, 2015). Abundance deregulation of particular signaling network nodes via genomic, transcriptional, or post-transcriptional regulatory defects (Feinberg, 2007; Bywater et al., 2013; Silvera et al., 2010) underlies human diseases, the prime example being cancer (Santarius et al., 2010). Copy number alterations of genes encoding critical proteins (Govindarajan et al., 2007; Eralp et al., 2008a; Han et al., 2015), independent of mutations that constitutively change enzymatic activity (Davies, 2002), drive progression of many cancer types. Genomic instability in cancer cells causes abnormally broad distributions of signaling protein abundances in a given tumor (Wang et al., 2015), yet the consequences of the protein abundance levels on signaling properties is poorly understood limiting our ability to rationally design therapies.

The epidermal growth factor receptor (EGFR) signaling network is affected by gene copy number alterations that deregulate protein abundances (e.g., of EGFR, HER2, ERK and AKT) in a number of cancer types (Govindarajan et al., 2007; Eralp et al., 2008b; Han et al., 2015). EGFR signaling controls cell growth, motility, survival, differentiation, and metabolism (Citri and Yarden, 2006). Many drugs target the activity of the EGFR signaling network (Tebbutt et al., 2013; Roberts and Der, 2007). The receptor tyrosine kinase (RTK) function of EGFR is activated by its dimerization upon ligand binding. EGFR auto-phosphorylation recruits adaptor proteins that typically activate the MAPK/ERK and AKT signaling pathways. The MAPK/ERK branch activates the GTPase RAS, which triggers a kinase phosphorylation cascade consisting of RAF, MEK, ERK, and p90RSK. The output of the MAPK/ERK branch is transcription of genes regulating growth and division (Mendoza et al., 2011; Olayioye et al., 2000). Signal transduction through the AKT branch starts by PI3K activation, producing PIP3, which recruits AKT and PDK1 to the plasma membrane. PDK1 phosphorylates AKT (Mendoza et al., 2011; Manning and Cantley, 2007), which mediates signaling through the mTORC1 complex to modulate

translation via p70S6K and 4EBP1 (Manning and Cantley, 2007). Other AKT targets are GSK3β, PRAS40, and TSC2. The AKT pathway controls cell survival, proliferation, and migration (Manning and Cantley, 2007). STAT proteins and the PKC pathway can also be activated by EGFR-mediated signaling (Bowman et al., 2000; Oliva et al., 2005). EGFR signaling involves crosstalk and feedback loops both internally (e.g., active ERK attenuates upstream RAF or MEK signaling via negative feedback) (Mendoza et al., 2011) and with other signaling pathways (e.g., WNT and TGF-β pathways) (Kim et al., 2007; Massague, 2003).

Classically, two approaches are used to characterize the effect of proteins on signal transduction. The first approach analyzes cell populations. Here, western blotting, mass spectrometry, RNA-microarrays, and synthetic lethality screens are used to identify signaling relationships (Zhang, 2005; Kim, 2012; Corcoran, 2013). Protein-protein interaction analyses are used to determine which proteins in a network directly interact (Kim, 2012; Tewari, 2004). Population-based methods yield a comprehensive view of signaling but are difficult to use in analysis of protein abundance dependencies due to inherent limitations: Proteins must be expressed at different abundances or cells must be sorted to yield a non-continuous abundance titration. Such methods result in a large number of samples and cell-to-cell protein abundance variations within each sample remain masked. The second approach studies signaling relationships in single cells. Here fluorescence microscopy and flow cytometry (FACS) are used with a variety of assays, including proximity ligation assay (PLA)(Sundqvist, 2013) or fluorescence resonance energy transfer (FRET)(Aoki, 2013). These approaches allow study of signaling relationships and dynamics through time and space; however, only a few signaling nodes can be measured simultaneously.

A recently developed single-cell analysis technology, called mass cytometry, allows for the simultaneous measurement of over 40 signaling nodes in single cells using metal-isotope tagged antibodies (Bodenmiller et al., 2012; Bendall, 2011). This capability makes mass cytometry uniquely suited to comprehensively query the function of nodes in signaling networks within heterogeneous cell populations. Mass cytometry is quantitative and, in

combination with mass-tag cellular barcoding (MCB), a powerful screening tool (Bodenmiller et al., 2012). Algorithms to analyze multiplexed single-cell mass cytometry data allow quantification of signaling relationships, therefore helping to decipher the highly complex network behaviors that operate even in simple biological systems (Krishnaswamy, 2014).

Here, we coupled protein overexpression with mass cytometry to measure the effect of varying node abundance on the activation state and signaling relationships of an unstimulated EGFR signaling network, as well as the signaling dynamics of the network in response to EGF stimulation. We exploited the finding that transient protein overexpression in a cell population typically produces a continuous abundance range of the target protein over four orders of magnitude. We overexpressed 20 central EGFR signaling network proteins individually in human embryonic kidney (HEK) 293T cells, sampled during an EGF stimulation time course over 60 minutes totaling 360 conditions. An average of 11,000 cells per condition was analyzed with a panel of 35 antibodies to provide a comprehensive single-cell proteomic EGFR network analysis. To identify signaling relationships in this dataset, we developed a statistical measure that we call 'binned pseudo R-squared' (BP-R^2) that recapitulated known signaling relationships and identified relationships that were to the best of our knowledge - not described previously. Thus, our experimental and computational approach enables study of how the strength and dynamics of signal transduction are tuned by node abundances.

2.3 Results

2.3.1 Analyzing continuous protein abundance dependencies

To systematically identify and characterize protein abundance-dependent signaling relationships, dynamics, and network activation states, we exploited the variation and large dynamic range of protein abundance induced by transient transfection and used mass cytometry to quantify the abundance of the transfected protein of interest (POI) in conjunction with comprehensive signaling network readouts in single cells. We cloned POIs genes into vectors containing a cytomegalovirus (CMV) promoter and a GFP-tag sequence (Couzens et al., 2013) to transiently overexpress GFP-tagged POIs in HEK293T cells (Figure 2.1a). The

tagged protein abundance was measured by mass cytometry using an anti-GFP antibody (Figure 2.1a). Ordering the measured cells based on the GFP signal provided a continuous POI titration (Figure 2.1b). Typically, not all cells were transfected, yielding an internal control for every experiment. To measure the single-cell EGFR signaling network states, we designed and validated a panel of 35 antibodies that mostly detect phosphorylation sites on signaling proteins (Supplementary Tables 1-3 online in Lun et al. (2017)). These data were used to determine the abundance dependencies of network activation state and signaling dynamics (Figure 2.1b).

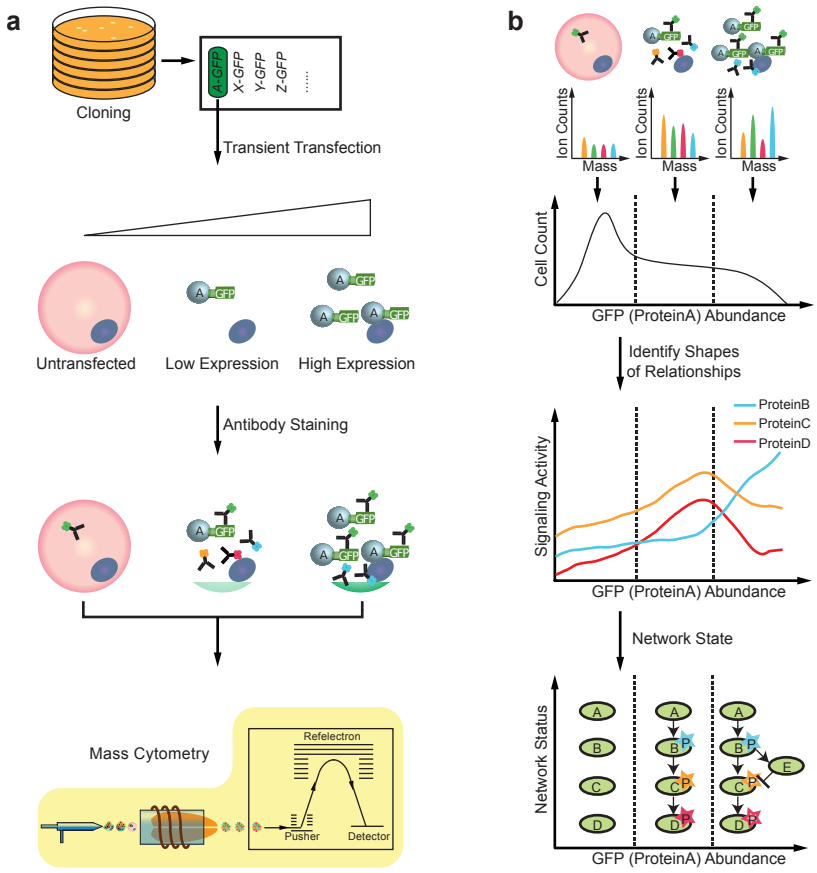


Figure 2.1: Workflow of abundance-dependent network analysis. (a) Experimental workflow. Signaling POIs are cloned into vectors containing a CMV promoter and a GFP-tag sequence to transiently overexpress GFP-tagged POIs in HEK293T cells. We quantify anti-GFP antibody as readout of POI-GFP abundance, together with other 35 markers, by mass cytometry. **(b)** Data analysis workflow. Cells were ordered based on the GFP signal, providing a continuous POI titration, which was then coupled to other signaling markers to determine the abundance dependencies of network activation state and signaling dynamics in the network after transfection. The network in the illustration does not represent an actual biological example.

To validate our system we confirmed that, first, the GFP tag was reliably detected by mass cytometry (Figure A.1); second, the GFP tag did not affect the localization and activity of the POI (Figures A.2, A.3; Supplementary Table 4 and Supplementary File 1 online in Lun et al. (2017)); third, POI expression levels were linearly related to GFP abundance, validating GFP as readout of the total POI abundance (Figure A.4a, c); fourth, POI overexpression for 18 hours (i.e., the time point of our experiments) did not alter the underlying network structure (Figure A.4b, c); fifth, the antibody-based GFP quantification by mass cytometry was comparable to FACS (Figure A.5); sixth, the cell culture media and cell detachment did not alter signaling processing in the EGFR network (Figures A.6, A.7); and, seventh, the levels of the GFP-tagged POIs were stable during the 1-hour EGF stimulation time course (Figure A.8, Supplementary Video 1 online in Lun et al. (2017)). We also found that the method is robust and highly reproducible as evidenced by the high concordance between the three individual experiment replicates (Figure A.9, Supplementary File 2 online in Lun et al. (2017)).

2.3.2 $\mathbf{KRAS^{G12V}}$ and $\mathbf{MEK1^{DD}}$ abundance effect on signaling

We first studied a well-known signaling circuit: Constitutively active mutants of KRAS and MEK1 ($KRAS^{G12V}$ and $MEK1^{DD}$) lead to ERK phosphorylation and activate components downstream in the MAPK/ERK pathway. As expected, we found that overexpression of $KRAS^{G12V}$-GFP or $MEK1^{DD}$-GFP increased phosphorylation on Thr202 and Tyr204 of ERK1/2 (Figure 2.2a). Our approach also elucidated the abundance-dependent effects on these signaling relationships: The relationship between $KRAS^{G12V}$-GFP and p-ERK1/2 was bow-like as high levels of $KRAS^{G12V}$-GFP corresponded to reduced phosphorylation of ERK1/2. By contrast, the $MEK1^{DD}$-GFP abundance relationship with p-ERK1/2 was monotonic as p-ERK1/2 increased with $MEK1^{DD}$-GFP expression (Figure 2.2a). These results verified the oncogenic activation of p-ERK1/2 induced by $KRAS^{G12V}$ and $MEK1^{DD}$.

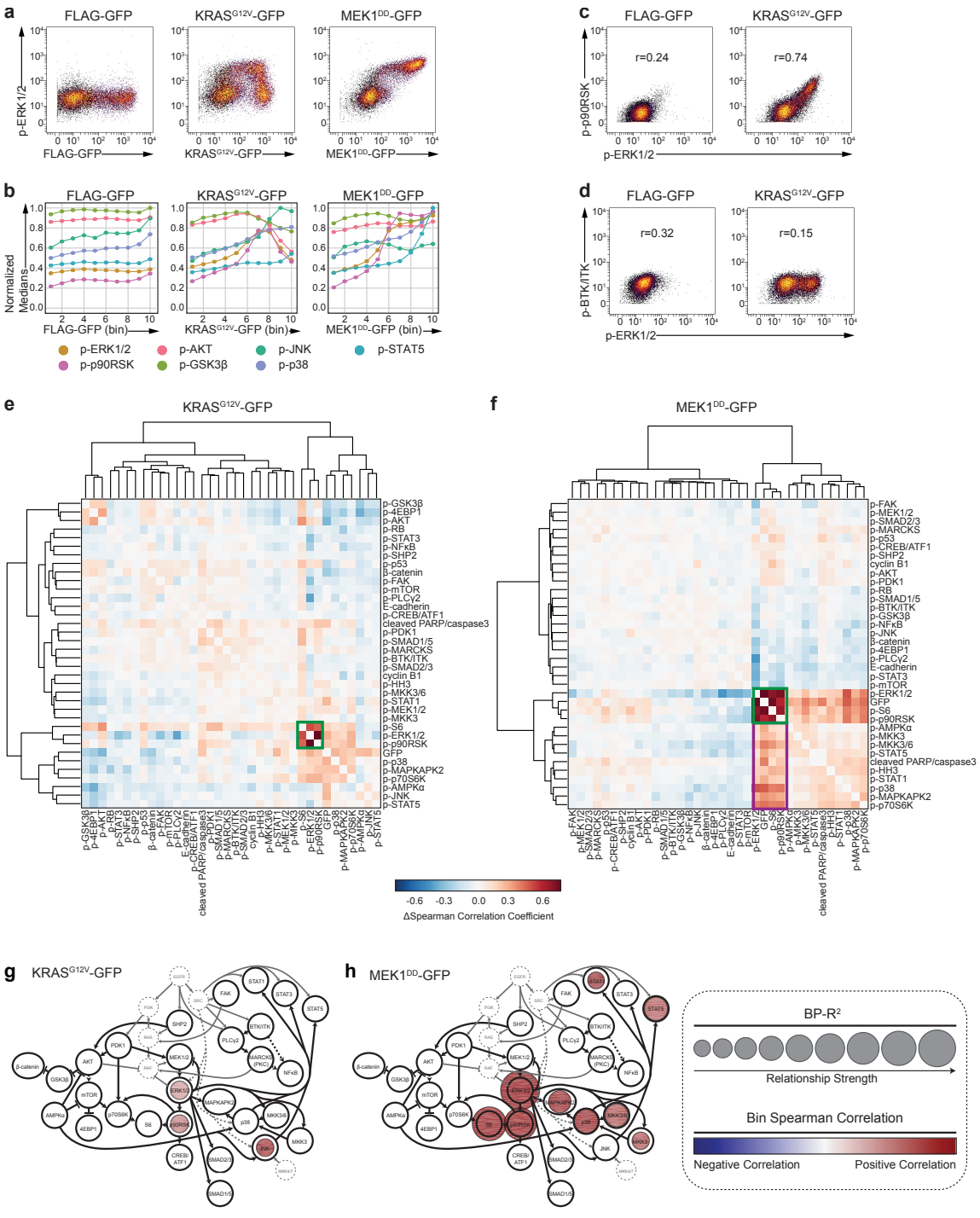

a
FLAG-GFP
KRASG12V-GFP
MEK1DD-GFP
p-ERK1/2
b
Normalized Medians
FLAG-GFP (bin)
KRASG12V-GFP (bin)
MEK1DD-GFP (bin)
p-ERK1/2
p-p90RSK
p-AKT
p-GSK3β
p-JNK
p-p38
p-STAT5
c
r=0.24
r=0.74
p-p90RSK
p-ERK1/2
d
r=0.32
r=0.15
p-BTK/ITK
e
KRASG12V-GFP
f
MEK1DD-GFP
ΔSpearman Correlation Coefficient
g
KRASG12V-GFP
h
MEK1DD-GFP
BP-R²
Relationship Strength
Bin Spearman Correlation
Negative Correlation
Positive Correlation

Figure 2.2: MAPK/ERK pathway mutants induce oncogenic signaling. (a) Biaxial plots of GFP, representing the abundance of the overexpressed mutant POIs, versus abundance of phosphorylation on Thr202/Tyr204 on ERK1/2. Constitutively active $KRAS^{G12V}$-GFP shows a downregulation on Thr202/Tyr204 on ERK1/2 at the highest levels of $KRAS^{G12V}$-GFP. Constitutively active $MEK1^{DD}$-GFP directly phosphorylates Thr202/Tyr204 on ERK1/2, and the abundance of the POI-GFP is correlated with amount of ERK1/2 phosphorylated at these sites. The FLAG-GFP control does not affect ERK phosphorylation sites. **(b)** The abundances of measured phosphorylation sites are plotted over the range of the $KRAS^{G12V}$-GFP and $MEK1^{DD}$-GFP expression. Phosphorylation sites of the same pathway (e.g., on ERK1/2 and p90RSK, AKT and GSK3, or p38 and JNK) show similar trends. An individual experiment is shown here. Plots for 3 replicates are shown in Figure A.9b-e. **(c)** Strong single-cell correlations within biaxial plots indicate co-regulated phosphorylation sites. **(d)** Unchanged and reduced correlations indicate unrelated phosphorylation sites. **(e)** and **(f)** Heat maps showing for all pairs of measured markers the change in Fisher-transformed Spearman correlation values for overexpression of **(e)** $KRAS^{G12V}$-GFP and **(f)** $MEK1^{DD}$-GFP when compared to the FLAG-GFP overexpression control. **(g)** and **(h)** BP-R^2 scores and Spearman correlations of bin medians for all measured markers in cells where **(g)** $KRAS^{G12V}$-GFP or **(h)** $MEK1^{DD}$-GFP was overexpressed overlaid on a literature-based graph of canonical signaling pathways (Roberts and Der, 2007; Mendoza et al., 2011; Massague, 2003; Kim, 2012; Cardaci et al., 2012; Rawlings et al., 2004; Xu et al., 2014; Nyati et al., 2006; Mitra et al., 2005; Hendriks et al., 2014). Strong relationships identified from the BP-R^2 analysis are plotted on the signaling maps as colored circles. The sizes of circles indicate relationship strengths quantified by BP-R^2. The directionalities of relationships, as judged by Spearman correlation of bin medians, are shown by the color of the circles (positive correlation indicates that cells show generally increasing marker levels, and a negative correlation indicates decreasing marker levels as POI-GFP levels increase). For (e) to (h), data from 3 individual experiment replicates were used.

Next, we analyzed the impact of $KRAS^{G12V}$-GFP and $MEK1^{DD}$-GFP abundance on all measured phosphorylation sites. We divided the measured cells into 10 bins according to the GFP signals and plotted the bin medians (Figure 2.2b, Figure A.9b-e). This analysis revealed that the phosphorylation site abundances on ERK1/2 and its direct downstream target Ser380 of p90RSK had similar relationships to the abundances of $KRAS^{G12V}$-GFP or $MEK1^{DD}$-GFP. Phosphorylation of AKT on Ser473 and its direct target Ser9 of GSK3β also had parallel trends and showed reduced levels when the MAPK/ERK signal peaked, suggesting inter-pathway regulation. We also observed increased JNK phosphorylation on Thr183/Tyr185 induced by the $KRAS^{G12V}$ mutant (Figure 2.2b) as reported previously (Zhou, 2010). This shows that our approach recapitulates known signaling relationships and identifies abundance-determined signaling responses.

We then systematically evaluated signaling relationships between all pairs of measured markers modulated by $KRAS^{G12V}$-GFP or $MEK1^{DD}$-GFP overexpression. We exploited the fact that overexpression of one protein increases signaling (i.e., phosphorylation levels) and thus expands the dynamic range of many measured markers (Figure 2.2c). This enabled the use of correlation analysis to distinguish signaling relationships (high correlation) from biological and technical noise (low correlation). For example, overexpression of $KRAS^{G12V}$-GFP resulted in an increased Spearman correlation between p-ERK1/2 and p-p90RSK compared to control (Figure 2.2c), whereas ERK-independent phosphorylation sites, such as Tyr551 of BTK/ITK, showed low correlation with p-ERK1/2 levels in both control and overexpression conditions (Figure 2.2d).

Identifying changes in pairwise Spearman correlations for all measured markers in the $KRAS^{G12V}$-GFP and $MEK1^{DD}$-GFP overexpression data compared to the FLAG-GFP control enabled systematic analysis of signaling relationship patterns (Figure 2.2e, f). Phosphorylation levels of proteins in the MAPK/ERK pathways showed strong increases in correlation, and pathway members clustered together (Figure 2.2e, f, green squares). We also observed that phosphorylations of MAPK/p38 pathway members and STAT proteins (STAT1 and STAT5) were increasingly correlated with levels of MAPK/ERK pathway members as MEK^{DD}-GFP levels increased (Figure 2.2f, purple rectangle), indicating crosstalk between MAPK and STAT pathways. These results reveal relationships among many measured markers and show that increases in correlation reflect pathways and grouped biological processes.

2.3.3 Automated analysis of abundance-induced signaling

Spearman correlation analysis can uncover strictly monotonic relationships between phosphorylation levels on signaling proteins; however, protein abundance-dependent signaling responses can be complex (Figure 2.2a, see $KRAS^{G12V}$). We therefore developed a density-independent measure termed 'binned pseudo R-squared' (BP-R^2) to quantify the strengths of relationships between the abundance of a POI and measured phosphorylation sites. BP-R^2 creates 10 bins across the POI-GFP expression range and calculates the relationship

strength considering bin medians and the global mean (Figure 2.3a, b, Methods, Supplementary Software online in Lun et al. (2017)). Using the BP-R^2 values for all negative controls, a cutoff for strong signaling relationships was determined (Figure 2.3c). Benchmarking BP-R^2 in identifying strong signaling relationships from the overexpression datasets showed that BP-R^2 outperformed methods often used for this task (Krishnaswamy, 2014; Redell, 2013) (Figure A.10a, b). The strong relationships identified by BP-R^2 were plotted in a two-dimensional layout guided by canonical pathways (Figure 2.2g, h). The directionality of measured signaling relationships was determined by Spearman correlation of the bin medians (Figure 2.3b, Methods). A positive correlation indicates that cells show generally increasing marker levels and a negative correlation indicates generally decreasing marker levels as POI-GFP levels increase.

Analysis of $KRAS^{G12V}$-GFP and $MEK1^{DD}$-GFP overexpression versus all measured markers using BP-R^2 revealed strong, positively correlated relationships of MEK^{DD}-GFP to downstream MAPK/ERK pathway nodes. $KRAS^{G12V}$-GFP levels, although also positively correlated with MAPK/ERK nodes, exhibited the same, but weaker relationships (Figure 2.2a, b, g, h). Together, these results suggest that feedback regulation of upstream MAPK nodes differs between the studied mutants. Additionally, this network view revealed that $MEK1^{DD}$-GFP abundance had a strong positive impact on nodes in the MAPK/p38 pathway; the previously observed $KRAS^{G12V}$-induced phosphorylation of JNK (Zhou, 2010) was dependent on $KRAS^{G12V}$ abundance (Figure 2.2g, h). These results show that overexpression of signaling proteins, in conjunction with BP-R^2 and correlation analysis, identifies known relationships and is a valid platform for discovery of signaling relationships in a comprehensive and abundance-dependent manner.

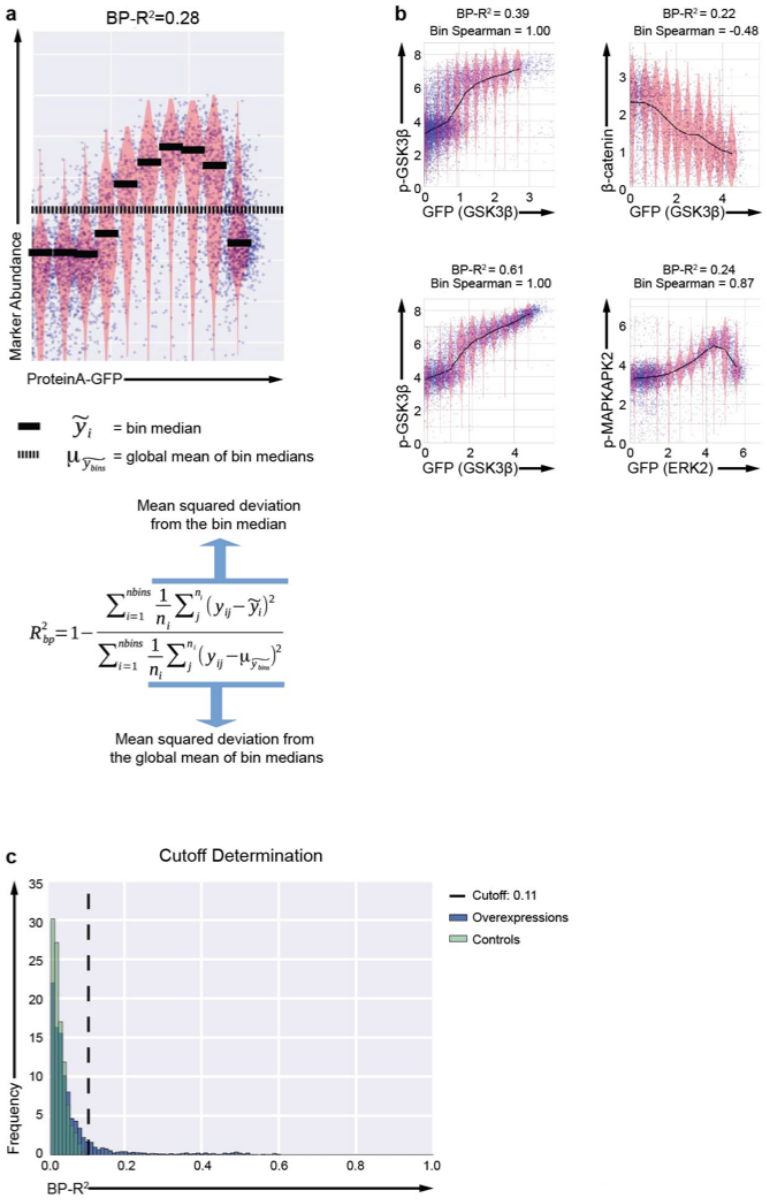
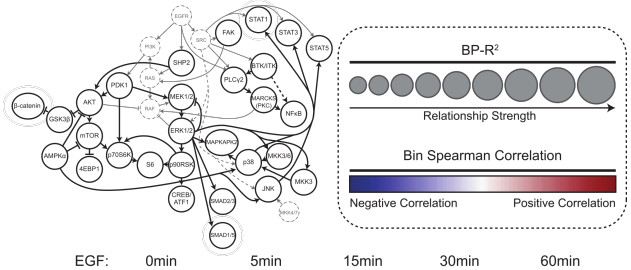


Figure 2.3: Binned pseudo R^2 (BP-R^2) analysis. **(a)** BP-R^2 analysis considers deviation from bin median versus the global mean of bin medians. **(b)** Examples of BP-R^2 and Spearman correlation of bin medians values. The top left and top right plots show examples of positive and negative Spearman correlations of bin medians. The top left and bottom left plots show replicates of the same overexpression condition and how a (supposedly) increased noisiness affects the BP-R^2 values. The bottom right plot shows a complex signaling relationship with the corresponding BP-R2 value. The BP-R^2 metric detects complex arbitrary relationship (bottom right). **(c)** Density distribution of the median BP-R^2 values for the 700 POI-GFP-marker relationships from the negative controls (FLAG-GFP, untransfected) and the 3500 POI-GFP-marker relationships of the signaling node overexpression conditions. Cutoff for strong signaling relationships were determined at a median BP-R^2 value of 0.11, the highest median BP-R^2 of the negative controls.

2.3.4 Node abundance dependency analyses of the EGFR network

To study the node abundance dependency of signaling relationships and dynamics in the EGFR signaling network, we overexpressed 20 EGFR-related signaling proteins individually in HEK293T cells (Table 2.1). Each of the 20 GFP-tagged POIs was validated in previous studies (Supplementary Table 5 online in Lun et al. (2017)) and in our system (Figures A.2 and A.3, Supplementary File 1 online in Lun et al. (2017)). 18 hours after transfection, we treated cells with EGF and quantified signaling by mass cytometry over a 60-min time course. To exclude signaling relationships caused by channel-to-channel spillover, we applied a stringent experimental filter (Figure A.11, Methods). The median marker intensities during the time course are shown in Figure A.12a. Based on these data we performed two sets of analyses. In the first, we used BP-R^2 analysis and Spearman correlations to evaluate how the abundance of overexpressed proteins influenced phosphorylation at the measured sites (Figure 2.4, Figure A.12b and Supplementary Files 2-4 online in Lun et al. (2017)). In the second, we examined how features of signaling dynamics depend on protein abundance (Figure 2.5).

In the first analysis, strong and broad signaling responses to overexpression were identified for the upstream kinases PDK1-, GSK3β-, SRC-, and ASK1-GFP without EGF stimulation (Figure 2.4. Overall, we identified 59 strong signaling relationships in the unstimulated conditions. Overexpression of many kinases induced strong and positively correlated signaling relationships with their own phosphorylation (Figure 2.4, Supplementary File 4 online in Lun et al. (2017)). Overexpression of CRAF-, KRAS-, p70S6K-GFP, and others only induced signaling responses upon EGF stimulation (Figure 2.4). Notably, under stimulated conditions, KRAS-, CRAF-, and MEK1-GFP levels negatively correlated with phosphorylation levels of downstream kinases p-ERK1/2 and p-p90RSK (Figure 2.4). Activating mutations in KRAS and CRAF (Figure 2.2), but not protein overexpression alone, may activate oncogenic signaling.


EGFR
PI3K
SRC
FAK
STAT1
STAT3
STAT5
SHP2
RAS
PDK1
PLCγ2
BTK/ITK
AKT
RAF
MEK1/2
AMOK3
(PKC)
NFκB
β-catenin
GSK3β
ERK1/2
mTOR
MAPKAPK2
AMPKα
4EBP1
p70S6K
S6
p90RSK
p38
MEKK3
JNK
MKK3
CREB/
ATF1
SMAD2/3
MEKK7
SMAD1/5
BP-R²
Relationship Strength
Bin Spearman Correlation
Negative Correlation
Positive Correlation

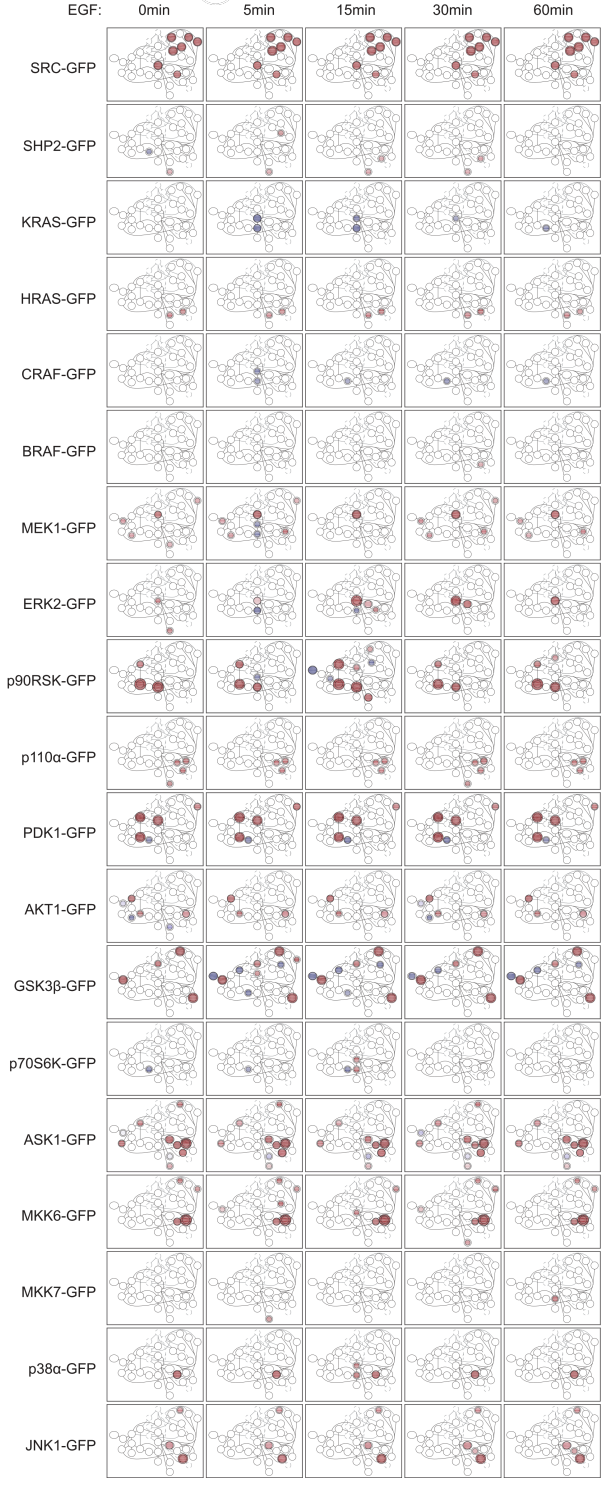

EGF:
0min
5min
15min
30min
60min
SRC-GFP
SHP2-GFP
KRAS-GFP
HRAS-GFP
CRAF-GFP
BRAF-GFP
MEK1-GFP
ERK2-GFP
p90RSK-GFP
p110α-GFP
PDK1-GFP
AKT1-GFP
GSK3β-GFP
p70S6K-GFP
ASK1-GFP
MKK6-GFP
MKK7-GFP
p38α-GFP
JNK1-GFP

Figure 2.4: Analysis of dynamics of EGFR signaling. HEK293T cells overexpressing GFP-tagged signaling proteins listed in Table 2.1 were treated with EGF for 0, 5, 15, 30, and 60 min. Strong abundance-dependent signaling relationships (Figure 2.3c) are plotted on the signaling map with circle sizes and colors indicating strengths (BP-R^2 score) and directionalities (Spearman correlation of bin medians), respectively. The miniaturized network is the same as used in Figure 2.2. Overexpression of S6-GFP did not induce any strong signaling relationships (data not shown). For all analyses, data from 3 individual experiment replicates were used.

Table 2.1: Overexpressed signaling proteins.

Overexpressed proteins	Gene ID	UniProt Entry
SRC	SRC	P12931
PDK1	PDPK1	O15530
AKT1	AKT1	P31749
GSK3β	GSK3B	P49841
MKK7	MAP2K7	O14733
MKK6	MAP2K6	P52564
p38α	MAPK14	Q16539
ERK2	MAPK1	P28482
p90RSK	RPS6KA1	Q15418
CRAF	RAF1	P04049
JNK1	MAPK8	P45983
p110α	PIK3CA	P42336
BRAF	BRAF	P15056
ASK1	MAP3K5	Q99683
p70S6K	RPS6KB1	P23443
MEK1	MAP2K1	Q02750
KRAS	KRAS	P01116
HRAS	HRAS	P01112
SHP2	PTPN11	Q06124
S6	RPS6	P62753

To systematically assess signaling relationships identified by BP-R^2, we used the literature curated signaling network, SIGNOR (Perfetto, 2016). For each relationship, we computed the shortest signed directed path length according to the SIGNOR network (Supplementary Table 6 online in Lun et al. (2017)). We found that 76% of the strong

relationships identified in the unstimulated conditions had paths with a maximum of three steps, highlighting that our approach identifies rather direct signaling relationships. Only 14 abundance-dependent relationships with four or more path steps were identified. Comparison of our strong signaling relationships with literature indicated that many EGF signaling connections that we identified were previously reported. We also propose many relationships that have – to our knowledge – not been previously reported, for example: p90RSK to PDK1 (Ser241), GSK3β to SHP2 (Tyr580), JNK1 to MAPKAPK2 (Thr334), p110 to MKK3 (Ser189), p110α to MKK6 (Ser207), ASK1 to PDK1 (Ser241), ASK1 to GSK3β (Ser9), and ASK1 to AMPKα (Thr172) (Table 2.2).

Table 2.2: Relationships with shortest singed directed path length above 3 in the SIGNOR database.

Overexpressed POI	Target	Sign	Shortest Signed Directed Path (SIGNOR)	Literature Information
SRC	p-BTK/ITK	1	6	SRC family kinases phosphorylate BTK (Hendriks et al., 2014)
SHP2	p-S6	-1	5	Known regulation (Marin, 2008)
ASK1	p-PDK1	1	5	Potential novel relationship
SRC	p-PLCγ2	1	5	SRC family kinases activates PLCγ2 (Hendriks et al., 2014)
ASK1	p-AMPK	1	4	Potential novel relationship
GSK3β	p-SHP2	1	4	Potential novel relationship
p90RSK	p-PDK1	1	4	Potential novel relationship
JNK1	p-STAT1	1	4	JNK activates STAT1 (Wei, 2014)
JNK1	p-MAPKAPK2	1	4	Potential novel relationship
p110α	p-MKK3/6	1	4	Potential novel relationship
HRAS	p-SMAD2/3	1	4	Known crosstalk (Massague, 2003)
ASK1	p-GSK3β	1	4	Potential novel relationship
PDK1	p-S6	-1	4	Overexpression-induced negative regulation
p70S6K	p-S6	-1	4	Overexpression-induced negative regulation

Phosphorylation levels of many members of the MAPK/ERK pathways showed complex relationships (i.e., measured phosphorylation levels varied over the analyzed POI-GFP range and the relationships did not fit linear, sigmoidal, or quadratic models) with levels

of POI-GFPs upon EGF stimulation. These relationships can be explained by abundance-dependent modulation of the signaling dynamics in response to EGF. Thus, in the second set of analyses we examined how signaling dynamics, as quantified by amplitude and peak-time, depended on abundance of an overexpressed protein (Figure 2.5). In order to view signaling trajectories as functions of protein abundance, we binned the POI-GFP levels into 10 bins (Figure 2.5a, Supplementary File 2 online in Lun et al. (2017)). This allowed tracing the signaling trajectories of cells with similar protein overexpression levels (i.e., those in the same bin) over the EGF stimulation time course (Figure 2.5b, Supplementary File 5 online in Lun et al. (2017)). Strong and robust changes in signaling amplitudes (Figure 2.5c-i) and peak-times (Figure A.13) were found. Notably, the maximum amplitudes were independent of the overexpression range of a given POI (Figure A.14).

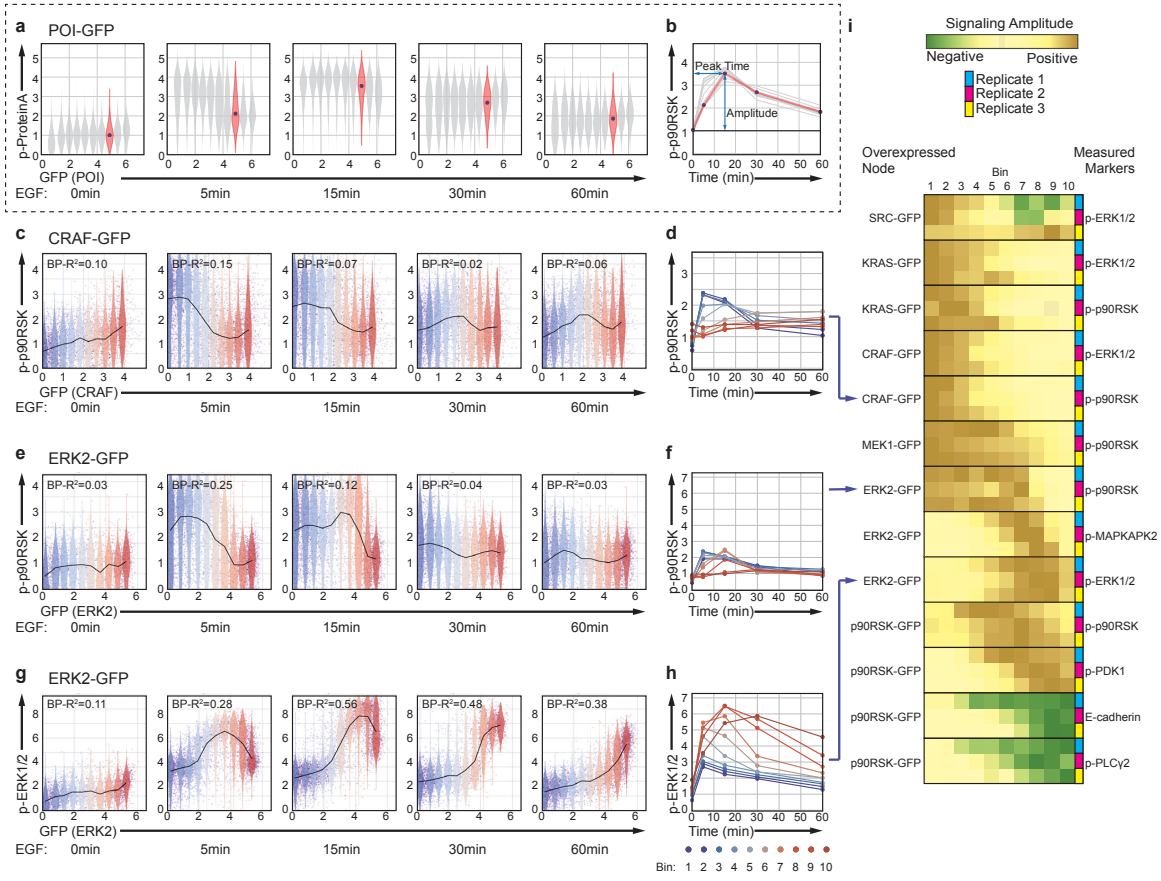


Figure 2.5: Analysis of node abundance-dependent EGFR signaling dynamics. **(a, b)** Schematic plots of amplitude and peak-time analysis. **(a)** The x-axis (i.e., overexpressed protein as determined by the GFP measurement) was split into 10 bins. **(b)** Median phosphorylation abundance in each bin was plotted on the y-axis versus time (x-axis) to visualize abundance dependency of signaling dynamics. **(c, d)** Mass cytometry ion counts (arcsinh transformed, Methods) measured for p-p90RSK (y-axis) as a function of ion counts measured for abundance of CRAF-GFP (x-axis) and EGF stimulation time. The same layouts for **(e, f)** ERK2-GFP abundance-determined p-p90RSK levels and **(g, h)** p-ERK1/2 levels are shown. **(i)** Heat map showing protein abundances with strong influences on signaling amplitudes with color indicating normalized signaling amplitudes. Only overexpressed proteins with an amplitude-ratio higher than 3 fold for more than two of the three replicates were identified as strong influences and are included in the heat map. For (a) to (h), representative examples from the 3 individual experiment replicates are shown. Other replicates are presented in Supplementary File 5 (online in Lun et al. (2017)). In (i), all replicate data are shown.

We found that high CRAF-GFP and KRAS-GFP abundance strongly reduced signaling amplitudes of p-ERK1/2 and p-p90RSK (Figure 2.5c, d, i), whereas high abundance of MEK1-GFP strongly reduced amplitudes and delayed peak-times for p-p90RSK (Figure 2.5i, Figure A.13). Overexpression of ERK2-GFP led to complex abundance-dependent responses of p-p90RSK and p-ERK1/2 after EGF stimulation (Figure 2.5e-h). p-ERK1/2 amplitudes increased and peak-times delayed as a function of ERK2-GFP abundance level (Figure 2.5g-i, Figure A.13). Intermediate abundance levels of ERK2-GFP also delayed the p-p90RSK peak-times relative to low ERK2-GFP abundance, whereas cells with high ERK2-GFP levels exhibited minimal p-p90RSK signaling dynamics (Figure 2.5e, f, i, Figure A.13). Overexpression of p90RSK-GFP modulated the signaling amplitude of its potential crosstalk phosphorylation site, Ser241 of PDK1, in an abundance-dependent manner, and increasing expression of p90RSK increased p-PDK1 amplitudes (Figure 2.5i). Thus, we observed abundance-dependent signaling dynamics across the range of overexpression levels. Overexpression of upstream signaling proteins (KRAS-, CRAF-, MEK1-, and ERK2-GFP) in the MAPK/ERK pathway led to reduced signaling amplitudes and delayed peak-times of their downstream targets. These observations show that our approach can quantify the role of protein abundance in determining the dynamic signaling response to an extracellular stimulation.

2.4 Discussion

Here we present an approach coupling transient overexpression with mass cytometry-based single-cell measurements to characterize signaling network activation states and signaling dynamics over a quasi-continuous, high dynamic range of protein abundance. To highlight the utility of our approach, we present a comprehensive single-cell proteomic analysis of the EGFR network that enabled an analysis of abundance-dependent effects of signaling proteins on state and dynamics of the signaling network. We evaluated the effects of overexpressing 20 EGFR network key nodes with a 60-minute EGF stimulation time course. In each of the 360 conditions, we measured the effect of a POI over a four order of magnitude abundance range on 35 markers by mass cytometry providing a unique and valuable quantitative single-cell resource of abundance dependencies of EGFR signaling.

Previously, the heterogeneity of protein levels after transient transfection was considered problematic. Here, we took advantage of this cell-to-cell variation as it results in a continuous titration of protein abundance over four orders of magnitude. Untransfected cells also provided an internal control for each experiment. We used the multiplexing capabilities of mass cytometry to characterize abundance dependencies of signaling network state and dynamics. Applied to the EGFR signaling network, our approach recapitulated known relationships, suggested previously not described ones, and revealed the intricate modulation of signal amplitudes and peak-times as functions of continuous protein abundance.

Our approach contributes to the understanding of signaling on several levels. First, the approach can be used to study uncharacterized proteins and to suggest additional roles to characterized ones. Second, we were able to directly relate POI abundance with the comprehensive analysis of signaling dynamics in response to stimulation. Such analyses are necessary for understanding of differential signal processing in identical cell types and in disease states characterized by heterogeneity in protein expression such as cancer. Third, the overexpression yields a large dynamic range of signaling activity and can reveal signaling relationships masked by stochastic processes and technical noise under otherwise similar conditions, facilitating the computational analysis of signaling relationships. Fourth, we present a metric termed BP-R^2, which allows the quantification of the strengths of arbitrary

shaped signaling relationships. BP-R^2 was superior to state-of-the-art methods for analysis of our dataset. Fifth, and finally, we were able to infer protein abundance-dependent signaling kinetics from single-cell snapshot data.

Our approach recapitulated known oncogenic signaling behaviors induced by the constitutively active mutants $KRAS^{G12V}$ and $MEK1^{DD}$ and identified novel abundance-dependent signaling relationships. For example, p-ERK1/2 was attenuated in cells with highly overexpressed $KRAS^{G12V}$-GFP, potentially due to negative feedback loops or senescence (Xu et al., 2014). Overexpression of the wild-type KRAS-GFP and MEK1-GFP did not induce downstream signaling activation, suggesting that mutations on KRAS or MEK1 are the main drivers of oncogenic signaling. Further, our approach allows study of abundance-dependent signaling dynamics. In the MAPK/ERK pathway, high abundance of upstream signaling mediators KRAS, CRAF, MEK1, or ERK2 reduced amplitudes and delayed peak-times of downstream phosphorylation sites. One possible explanation is that the signal transduction is determined by the competition between active and inactive forms of a signaling protein for substrates. Overexpression increases the total abundance but may reduce the percentage of the active form.

KRAS amplification has been identified in many cancer types. Amplification, however, is not correlated with the phosphorylation of ERK1/2 (Rahman, 2013). Rather, *KRAS* amplification mediates resistance to inhibitors targeting growth pathway related kinases, including EGFR, MET and MEK1/2; KRAS knockdown diminishes the drug resistance (Valtorta, 2013; Cepero, 2010; Little, 2011). Our results indicate that due to reduced downstream signaling amplitudes in response to EGF stimulation, the dependency of cells on the MAPK/ERK pathway may decrease upon KRAS overexpression, suggesting a mechanism for cancer cell resistance to inhibitors.

Comparing the identified strong signaling relationships with those in the SIGNOR database, we propose previously not described signaling relationships, e.g.: 1) Our data suggest that p90RSK potentially forms a positive feedback loop and activates the upstream signaling protein PDK1. 2) GSK3β has been identified as a central signaling controller and

has multiple substrates (Cohen and Frame, 2001); our results suggest that SHP2 is a potential direct or indirect target of GSK3β. 3) We also propose that JNK1 is a MAPKAPK2 activator. 4) PI3K and MKK3/6 are known to be regulated by RAC1 (Shin et al., 2005); our results suggest PI3K activates MKK3/6 independently. 5) Recent studies indicate that ASK1 contributes in negative regulation of PDK1 through phosphorylation on Thr254 of PDK1 (Seong et al., 2010); We observe ASK1 overexpression-induced PDK1 phosphorylation on Ser241, inducing PDK1 activity and downstream GSK3β phosphorylation on Ser9. 6) In addition to the known AMPK-mediated ASK1 activation (Lee et al., 2010), our data indicates ASK1 activation of AMPK via phosphorylation on Thr172. 7) We have also observed negative correlations between the abundance of p70S6K or PDK1 to the phosphorylation level of S6 (Ser235/Ser236), indicating overexpression-induced-negative feedback regulations.

Our method has several limitations. First, we do not measure the endogenous expression level of the POI. However, exogenous expression is linearly correlated with the total protein level (Figure A.4a), validating GFP as readout of the total POI. Second, all results in mass cytometry rely on antibodies; for this work, all antibodies were thoroughly validated (Supplementary Table 3 online in Lun et al. (2017)). Third, we do not measure the abundance range of the studied proteins in cancer cells, however, proteome studies of cancer cells and databases such as PaxDb (Wang et al., 2015) indicate a range similar to those studied here. Fourth, high expression levels of a protein kinase may induce non-specific phosphorylation; however, our data allows choosing the analyzed expression range in silico, thus such effects can be excluded.

The approach described here provides a method to study how the abundance variance of signaling proteins in different tissues and cell lines results in distinct signaling behaviors. The application of our approach to synthetic biology, stem cell biology, developmental biology, and cancer-related processes, such as the epithelial-mesenchymal transition, will enable quantitative identification of key proteins and signaling determinants in cell differentiation at phenotypical switching points. We envision that determining which signaling relationships and thresholds enable diseased cells to overcome drug treatment will be a

highly relevant application.

2.5 Methods

Data availability

All raw data are available at `http://www.cytobank.org/bodenmillerlab` and `http://www.bodenmillerlab.org/`.

Cloning

DNA sequences of the genes of interest were provided in entry clones by William Hahn and David Root (Yang et al., 2011) (via Addgene and NEXUS Personalized Health Technologies at ETH Zurich). Destination vectors, including pDEST pcDNA5 FRT TO-eGFP, pDEST 5' Triple Flag pCDNA5 FRT TO and pDEST 3' Triple Flag pcDNA5 FRT TO, were kindly provided by Dr. Anne-Claude Gingras at Lunenfeld-Tanenbaum Research Institute, Toronto, Canada (Couzens et al., 2013). Expression vectors encoding the FLAG- or GFP-tagged fusion proteins were generated via Gateway Cloning and sequenced before transfection.

Cell culture

HEK293T cells, obtained from ATCC, were cultured in DMEM (D5671, SIGMA), supplemented with 10% FBS, 2 mM L-glutamine, 100 U/ml penicillin, and 100 μg/ml streptomycin. For cell passaging or harvesting, cells were incubated with 1X TrypLE^{TM} Express (Life Technologies) for 2 minutes at 37°C.

Transfection and stimulation

HEK293T cells were seeded at the density of 0.7 million per well in 6-well plates. After 24 hours, cells were transfected with 2 μg plasmid and 4 μl of jetPRIME (PolyPlus) per well with the standard protocol provided by the manufacturer. At 18 hours after transfection, EGF (Peprotech) was added to a final concentration of 100 ng/ml. At 20 minutes before a given EGF stimulation time point, 5-Iodo-2-deoxyuridine (IdU) was added to the medium at the final concentration of 10 μM. At 2 minutes before a given EGF stimulation time point, medium was replaced by 1X TrypLE to induce cell detachment. At the time

point, paraformaldehyde (PFA, from Electron Microscopy Sciences) was added to the cell suspension to a final percentage of 1.6%, and cells were incubated at room temperature for 10 minutes. If EGF stimulation was not necessary in the experiment, cells were directly harvested and crosslinked with PFA. Crosslinked cells were washed twice with cell staining media (CSM, PBS with 0.5% BSA, 0.02% NaN_3) and after centrifugation, ice-cold methanol was used to resuspend the cells, followed by a 10-minute permeabilization on ice or for long-term storage at -80°C. Three individual experiment replicates were performed for each experiment. In each replicate, the experimental procedures were performed on different days.

Cell sorting

HEK293T cells overexpressing FLAG-GFP were detached from the plates as described above and resuspended in the FACS buffer (PBS with 0.5% BSA and 5 mM EDTA). Cells were sorted with BD FACSAria III Cell Sorter into GFP low, intermediate, and high levels with the strategy indicated in Figure A.5.

Live cell imaging

HEK293T cells were seeded in CultureWell^{TM} Chambered Coverglass (Thermo Fisher Scientific) pre-coated with fibronectin. Transfection of FLAG-GFP was performed as described above. At 18 hours after transfection, cells were imaged with a Leica DMI 6000 inverted microscope at 37°C and 5% CO_2. Images were taken every 2 minutes for 1 hour.

Immunofluorescence staining

Cells were cultured in 16-well glass chamber slides (Thermo Fisher Scientific). Transfection was done as described above. Before staining, culture medium was removed, and the slide was then washed with PBS. To crosslink cells, 4% PFA was added, and cells were incubated at room temperature for 20 minutes. Slides were washed with PBS three times, and cells were permeabilized for 5 minutes with 0.1% TritonX-100 dissolved in PBS at room temperature. After washing three times with PBS, cells were incubated in blocking buffer (10% goat serum diluted in PBS) for 30 minutes at room temperature. Primary (anti-GFP, FM264G,

BioLegend, 1:200) and secondary (Goat anti-Rat Alexa Fluor® 568, 1:500, supplemented with Hoechst 33342 at a final concentration of 100 μg/ml) antibodies were diluted in blocking buffer and applied to slides. A total protein stain of Alexa Fluor 647 Succinimidyl Ester (Life Technologies) was used to indicate cell outlines. Cells were washed three times with PBS after each incubation step. Slides were mounted with ProLong Gold Antifade Reagent (Life Technologies) before imaging with a CLSM Leica TCS SP8 microscope.

Antibody conjugation

The MaxPAR antibody conjugation kit (Fluidigm) was used to generate isotope-labeled antibodies using the manufacturers standard protocol. After conjugation, the antibody yield was determined based on absorbance of 280 nm. Candor PBS Antibody Stabilization solution (Candor Bioscience GmbH) was used to dilute antibodies for long-term storage at 4°C.

Barcoding and staining protocol

Formalin-crosslinked and methanol-permeabilized cells were washed three times with CSM and once with PBS. Cells were incubated in PBS containing barcoding reagents (^{102}Pd, ^{104}Pd, ^{105}Pd, ^{106}Pd, ^{108}Pd, ^{110}Pd, ^{113}In and ^{115}In) at a final concentration of 100 nM for 30 minutes at room temperature and then washed three times with CSM (Bodenmiller et al., 2012). Barcoded cells were then pooled and stained with the metal-conjugated antibody mix (Supplementary Table 1 online in Lun et al. (2017)) at room temperature for 1 hour. The antibody mix was removed by washing cells three times with CSM and once with PBS. For DNA staining, iridium-containing intercalator (Fluidigm) diluted in PBS with 1.6% PFA was incubated with the cells at 4°C overnight. On the day of the measurement, the intercalator solution was removed, and cells were washed with CSM, PBS, and ddH_2O. After the last washing step, cells were resuspended in ddH_2O and filtered through a 70-μm strainer.

Mass cytometry analysis

EQ^{TM} Four Element Calibration Beads (Fluidigm) were added to cell suspensions in a 1:10 ratio (v/v). Samples were analyzed on a CyTOF2 (Fluidigm). The manufacturers standard operation procedures were used for acquisition at a cell rate of 500 cells per second. After the acquisition, all FCS files from the same barcoded sample were concatenated (Bodenmiller et al., 2012). Data were then normalized, and bead events were removed (Finck et al., 2013) before doublet removal and de-barcoding of cells into their corresponding wells using a doublet-filtering scheme and single-cell deconvolution algorithm (Zunder et al., 2015). Subsequently, data was processed using Cytobank (`http://www.cytobank.org/`). Additional gating on the DNA channels (^{191}Ir and ^{193}Ir) and ^{139}La/^{141}Pr was used to remove remained doublets, debris and contaminating particulates.

Data visualization and analysis

Bi-axis scatter plots

Bi-axis scatter plots were generated in Cytobank (http://www.cytobank.org/).

Data preprocessing

Raw data was transformed using the inverse hyperbolic sine transform with a cofactor of 5:

$$data = \text{arsinh}(data_{raw}/5)$$

Except where use of raw data values is specifically noted, all visualizations and analyses were performed using transformed data.

Data binning

For data binning, the range between the lower and upper 2.5% of observations was divided into ten equal bins: bin_1,...,bin_{10}. The observations in the lower and upper 2.5% were assigned to the lowest and highest bins, respectively. In order to be able to compare expression levels between samples within a time course replicate, all observations of the time course were used to determine the binning.

Correlation analysis

Spearman correlation (r_{ij}) was calculated between all marker pairs (i, j) for each replicate and condition. Fishers z-transformation:

$$z_{i,j} = \text{artanh}(r_{ij})$$

was used to compare pairwise correlation coefficients across conditions. For each overexpressed POI, the change in correlation matrix was calculated by subtracting the median $\tilde{z}_{ij}$ value (across replicates) of the FLAG-GFP controls from the median $\tilde{z}_{ij}$ value (across replicates) of the overexpression condition.

$$\Delta\tilde{z}_{ij} = \tilde{z}_{ij\ \text{overexpression}} - \tilde{z}_{ij\ \text{FLAG-GFP}}$$

The resulting matrix of differences in Fisher-transformed correlation values was hierarchically clustered using the Ward method and Euclidean distances (Ward, 1963).

BP-R^2

Relationships between overexpression levels and markers can include non-monotonic relationships that are not properly captured with correlation metrics such as Spearman correlation. Furthermore, although the shapes induced by an overexpression were highly reproducible, the number of cells with a given expression intensity level were not. Thus, in order to quantify the strength of arbitrarily shaped relationships between markers and overexpression levels over the whole overexpression range, a density agnostic metric termed binned pseudo R-squared (BP-R^2) was developed. For this metric, the middle 95% of POI-GFP levels over a time-course experiment was divided into 10 equal-width bins. Bins with less than 25 cells were discarded. For each bin i, the median of a measured marker ($\tilde{y}_i$) was calculated. Additionally, the overall mean of the medians of all the 10 bins ($\mu_{\tilde{y}}$) was calculated. Then, for each bin, we computed the sum of squared deviations from the bin medians and the sum of squared deviations from the overall mean of medians. These values were summed over all bins, and the BP-R^2 was defined as one minus the ratio between these

two values:

$$R^2_{\mathrm{BP}} = 1 - \frac{\sum_{i=1}^{nbins} \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \tilde{y}_i)^2}{\sum_{i=1}^{nbins} \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \mu_{\tilde{y}})^2}$$

Following the rationale of classical R-squared statistics, BP-R^2 quantifies the average reduction in squared deviations per bin when modeling the data as piecewise constant within each bin (based on the bin medians) compared to using the mean over all bin medians. The BP-R^2 metric represents the relationship strength between a marker and the overexpressed signaling protein relative to the overall variability of the marker. By using the median instead of mean, the BP-R^2 selects unimodal relationships with low noise over noisy, multimodal relationships. Notably this measure works with arbitrary interaction shapes and is largely robust against density inhomogeneities. In order to aggregate the sample replicates, we considered the median BP-R^2 value across the experimental triplicates.

Threshold determination

We observed many relationships with low BP-R^2 between overexpressed proteins and measured phosphorylation markers, even within control samples (overexpression of FLAG-GFP). We propose that such weak relationships are more likely to result from indirect biological mechanisms; therefore, we focused on relationships that were stronger than all relationships seen in the controls (FLAG-GFP overexpression and untransfected cells). We chose the maximum median (across replicates) BP-R^2 of all controls (FLAG-GFP overexpression and untransfected cells) as a cutoff. Relationships that had a median BP-R^2 higher than this threshold were considered as sufficiently strong to be of interest.

Kinetic analysis

For each overexpression condition, replicates of EGF stimulation time courses were processed, stained, and measured together. Simultaneous processing enabled direct quantitative comparisons of the measured POI-GFP counts in these time courses. Samples representing all time points in a time course replicate were combined and binned by POI-GFP intensity as described in the **Data binning** section. As the binning was performed over all

samples of the same time course, the range of GFP intensity of bins with the same bin index directly corresponds to cells with similar abundance levels of POI-GFP in each of the different time points. As POI-GFP levels stay quasi-constant over the timescale of the 60-minute time course (Figure A.8, Supplementary Movie 1 online in Lun et al. (2017)), tracking how the median marker levels in a specific bin change over the time course reflects the kinetics of cells with a similar abundance level upon stimulation. Thus, the kinetic responses over a range of cells with low-abundance to high-abundance POI can be compared and analyzed using classical signal processing readouts such as signal response amplitude and peak-time. For this analysis, only POI-GFP marker pairs with at least one strong relationship over the time course were considered.

Amplitude analysis

The response amplitude for each binned abundance level was calculated using raw counts. For each measured marker and time point, the median marker level of each POI-GFP bin was divided by the median level of the marker in the corresponding bin of the unstimulated sample (EGF 0 min) to calculate amplitude as a fold change. The amplitude for each bin was identified as the maximum fold change over all time points. Robust and strong abundance-dependent changes were identified by comparing the amplitude ratio between the second highest and the second lowest bin amplitude. Of those identified as robust and strong, overexpressed proteins with an amplitude ratio higher than 3 fold for at least 2 of the 3 replicates were identified as interesting and plotted as a heat map.

Peak-time analysis

Interesting examples of overexpression changes were defined by identifying the time point with maximum amplitude for each bin (referred to as the peak time). Consistent and robust examples were selected by the following criteria: Monotonically increasing or decreasing peak times with increasing POI and a clear change in amplitude (>3 fold) in at least 2 of the 3 replicates.

SIGNOR database comparison

To compare the consistency between the strong signaling relationships detected by BP-R^2 analysis with the relationships predicted by the SIGNOR database (Perfetto, 2016), the python NetworkX (Hagberg et al., 2008) package was used to construct a sign-directed SIGNOR network from the UniProt entries of overexpressed POIs and measured phosphoproteins. NetworkX also calculates a shortest path length within the sign-directed network. Antibodies may bind to the same phosphorylation sites on more than one protein from a family, making the mapping between antibodies and UniProt entries ambiguous. In this case, the shortest path value was calculated between the overexpressed POI and any possible antibody targets.

The analysis was performed including the directionalities of signaling relationships as identified by the Spearman correlation of the bin medians in our analysis, and by exploiting the SIGNOR annotations in the following way: Simple paths between the overexpressed protein and the targeted phosphorylation sites were analyzed, starting from the shortest path to longer paths, until a sign-consistent path was found. To identify sign consistency, all edges in the SIGNOR network were classified as positive, negative, or ambiguous based on the SIGNOR Effect annotation: down-regulates = negative, up-regulates = positive, something else = ambiguous. In cases for which there were multiple interaction types possible for an edge (positive and negative), the overall sign was taken to be ambiguous. In cases where the last edge was annotated to be affecting exactly the residue (SIGNOR annotation Residue) measured by the phospho-specific antibody through (de)phosphorylation (SIGNOR annotation Mechanism), the directionality sign of this edge was determined to be phosphorylation = positive, dephosphorylation = negative or the inverse in cases the antibody was measuring the non-phospho site (e.g., Ser33/37/Thr41 on β-catenin). Measured phosphorylation sites responsible for inactivating a protein (e.g., Ser9 on GSK3β) were also signed as phosphorylation = negative. A path was determined to be sign consistent if the product of the signs of all its edges was in accordance with the relationship direction as measured by the Spearman correlation over the bins.

Systematic spillover exclusion

A stringent spillover filter was applied to systematically remove strong signaling relationships potentially affected by channel-to-channel spillover: For any measured channel that had events with ion counts over 500, we checked for spillover due to, first, isotope impurity (channels with isotopes of the same metal); second, mass resolution (-1 and +1 channels); and, third, oxidation (+16 channels). Any strong relationships (BP-R^2) with GFP and markers from these sets of channels were selected for additional verification experiments, in which the staining was done in three groups:

1. All antibodies in a set
2. All antibodies in a set except for the one that potentially causes spillover
3. Only the antibody potentially causing spillover

When spillover-induced background contributed to over 10% of the actual ion counts, the channel was discarded from further analysis (Figure A.11).

Based on our spillover exclusion protocol, we found the following channels were affected by spillover. They were excluded from the analysis performed in this manuscript, and should not be considered in any subsequent analyses using this data:

In SRC overexpression:

^{142}Nd p-SHP2

^{159}Tb p-SMAD1/5

In PDK1 overexpression:

^{142}Nd p-SHP2

^{143}Nd p-FAK

^{145}Nd p-MAPKAPK2

^{160}Gd p-MKK3/MKK6

^{162}Dy p-BTK/ITK

In GSK3β overexpression:

^{146}Nd p-p70S6K

In p90RSK overexpression:

^{142}Nd p-SHP2

^{147}Sm p-MKK3

^{162}Dy p-BTK/ITK

^{164}Dy p-SMAD2/3

CHAPTER III

VARIATION IN SINGLE-CELL SIGNALING DYNAMICS IS DETERMINED BY INITIAL CELL STATE[1]

3.1 Abstract

Diseased cells display abnormal variations in protein expression and function. In particular, cell-to-cell differences in signaling components can lead to qualitatively different functional responses within a cell population, for instance, in cancer. Understanding the origins of this response heterogeneity is hampered by the inability of time-lapse methods to measure multiple pathway components simultaneously. Here we present a computational method to infer single-cell signaling trajectories from multiplexed snapshot data and use it to analyze signaling dynamics in the extracellular signal-regulated kinase (ERK) pathway. The results demonstrate that the pathway is tuned to transmit signal duration, rather than signal amplitude. We predict and subsequently confirm the existence of cells that are insensitive to a chemical inhibitor of ERK signaling. Finally, we use ERK overexpression to explore signaling rate functions within an abnormal disease expression range. Taken together, our results show cell-to-cell variation in MAPK/ERK signaling responses depends primarily on initial cell state.

3.2 Introduction

Cell signaling pathways are complex biochemical systems at the core of cellular information processing, and the dynamics of these signaling systems in response to extracellular cues plays a critical role for proper cell functioning (Dolmetsch et al., 1997; Kholodenko, 2006; Selimkhanov et al., 2014). Signals are transduced by modulating the enzymatic activities and local concentrations of signaling mediators such as protein kinases. Cell-to-cell variation in expression of signaling components, even within a clonal cell population, can lead to

[1]This chapter is adapted from a manuscript currently in revision after peer review at **Cell** and has the following author list: Wade, JD, Lun XK, Bodenmiller B, Voit EO.

different responses and functional outcomes, such as proliferation versus apoptosis (Spencer et al., 2009). In cancer and other diseases, genetic alterations change the expression or function of signaling components, increasing cellular variation and creating cells with aberrant signaling responses compared to healthy cells (Altschuler and Wu, 2010). In addition, microenvironmental differences can further increase mixtures of cancer cell subpopulations that qualitatively differ in their molecular expression profile (cell state) and, consequently, drug responses (Meacham and Morrison, 2013; Burrell and Swanton, 2014). In such cases, one cell subpopulation may strongly respond to a particular drug treatment while another is insensitive (Burrell and Swanton, 2014).

The MAPK/ERK signaling cascade controls multiple cellular decisions, including cell motility, growth, proliferation, differentiation and survival (Anjum and Blenis, 2008; Santos et al., 2007). Multiple receptor types can activate the cascade, where "signal" passes from RAF to MEK to ERK protein kinases through activating phosphorylations. Activated ERK then regulates transcription and translation by phosphorylating multiple transcription factors and downstream kinases, such as p90RSK (Caunt et al., 2015). The expression and/or function of ERK pathway components is altered in many cancers and the pathway is a common target of cancer drugs (Samatar and Poulikakos, 2014; Caunt et al., 2015). Due to variation in tumor cell states, however, targeted therapies have had more limited success than expected (Wellbrock and Arozarena, 2016; Caunt et al., 2015). Successful therapies, therefore, must account for variation in expression of signaling components and its effects on signaling dynamics at the single-cell-level.

To characterize cell-to-cell variation in signaling dynamics, the changes in abundances or, ideally, concentrations of signaling molecules within individual cells must be observed over time. This characterization is experimentally possible through live-cell imaging methods such as Förster resonance energy transfer (FRET), kinase translocation reporters (KTRs) and activity-dependent dyes that measure the activation states of signaling molecules (Ryu et al., 2015; Regot et al., 2014; Dolmetsch et al., 1997). However, cell signaling networks consist of numerous interacting components, whereas live-cell methods are currently limited

to observation of one pathway component at a time (Bunt and Wouters, 2017). As an alternative, multiplexed single-cell methods, such as fluorescent and mass cytometry, are able to characterize the state of signaling networks in a more comprehensive manner by simultaneously measuring up to 50 components (Lin et al., 2015; Giesen et al., 2014; Bodenmiller et al., 2012; Lun et al., 2017), but require fixation and can provide only a snapshot of single cell states. Expressed differently, such snapshot methods are unable to follow the signaling dynamics of an individual cell over time, thus losing the critical link between initial cell state and response to stimulus.

Computational methods have been used at the cell population-level to overcome the trade-off between continuous and snapshot measurements. In the case of snapshot measurements, a mathematical model of the underlying system is reverse-engineered and used to infer the characteristics of continuous trajectories. The underlying biochemical reactions of cell signaling networks are most commonly modeled by a system of deterministic ordinary differential equations (ODEs). The use of deterministic models in biochemical processes such as signaling is generally justified by the large abundance of reaction components, which smooths out stochastic variability in reactions (Wang et al., 2015; Filippi et al., 2016). Thus, forward-simulation of the ODE model generates continuous trajectories of state variables, such as the phosphorylation states of signaling proteins on a cell population level. Snapshots of simulated trajectories are then compared to experimental measurements to determine model quality and infer model parameters.

Population-level models, however, do not account for single-cell variations in a sample, despite their well-established importance. This failure to account for variation in signaling components is intrinsic and mandates the development of new approaches based on single-cell measurements (Bronstein et al., 2015); especially snapshot measurements, since these enable highly multiplexed observations. Current methods to model variation in signaling dynamics use parametric distributions to represent single-cell snapshot data (Hasenauer et al., 2011, 2014; Filippi et al., 2016), where model simulations describe trajectories of additional higher-order distributional moments, beyond the mean (e.g., variance), which describe features of the distribution shape and statistically quantify the appropriateness of

the model quality. Use of parametric distributions, however, complicates model inference by adding parameters that must be inferred from data. As the number of parameters rapidly increases with the number of state variables (e.g., signaling network components) and/or as samples contain increasingly complex mixtures of subpopulations (distributions), parametric problem formulations rapidly become computationally and/or statistically intractable. Currently, these problems are addressed by ignoring cross-moments (e.g., covariance) (Hasenauer et al., 2014; Filippi et al., 2016), although these feature characterize how variables co-depend and contain valuable information unique to single-cell data (Sachs et al., 2005). Furthermore, parametric methods do not account for single-cell trajectories. Due to these limitations and despite the increased ease of acquiring multiplexed single-cell datasets, parametric methods have only been used to model two-dimensional experimental (Hasenauer et al., 2014; Filippi et al., 2016) or simulated (Hasenauer et al., 2011) data.

Here we propose a novel ODE modeling approach to infer multiplexed single-cell signaling trajectories from snapshots of multiplexed single-cell time-course data that does not depend on parametric distributions. Our methodology uses an ODE system to simulate the trajectories of many individual cells, rather than distribution parameters, and applies a distribution-free, rather than parametric, statistical test to compare experimental and simulated snapshots and inform model fitness. This decouples the number of model parameters from the number of model variables or subpopulations. Additionally, our approach is straightforward to implement, fits easily within current modeling frameworks and quantifies latent (unmeasured) sources of variation in cell state that contribute to a signaling process and may subsequently be used to discover new biology.

We use our approach to study single-cell variation within the MAPK/ERK signaling cascade in HEK293T cells stimulated with epidermal growth factor (EGF). In addition to normal signaling responses, we analyze a condition with drug treatment and, as the progression of many cancers is related to protein overexpression, a disease state where protein expression has been altered. Ultimately, we show: that, given initial cell states, single-cell variation in signaling responses can be described deterministically; that, while each level of the MAPK/ERK pathway has a fold-change response independent of other factors, the

entire pathway is tuned to robustly transmit signal duration; that the kinetics for "disease" cell states cannot necessarily be inferred from "normal" cells; and, that our methodology may be used to find predictors of drug response. Our results reinforce the importance of considering signaling-dynamics at the single-cell level for understanding variation in cellular responses and illustrate the usefulness of our technique to do so.

3.3 Results

3.3.1 Inference of signaling trajectories from multiplexed single-cell snapshots

Determining sources of cell-to-cell variation in signaling responses requires two components: the cell state variables that determine signaling response, and single-cell signaling trajectories to characterize response. Unfortunately, current experimental methods cannot capture both simultaneously; for instance, measurements of single-cell trajectories generally lack the dimensionality to observe variables driving differential response (Figure 3.1a), while higher-dimensional snapshot measurements lack the time-dependent single-cell dynamics needed to quantify response (Figure 3.1b).

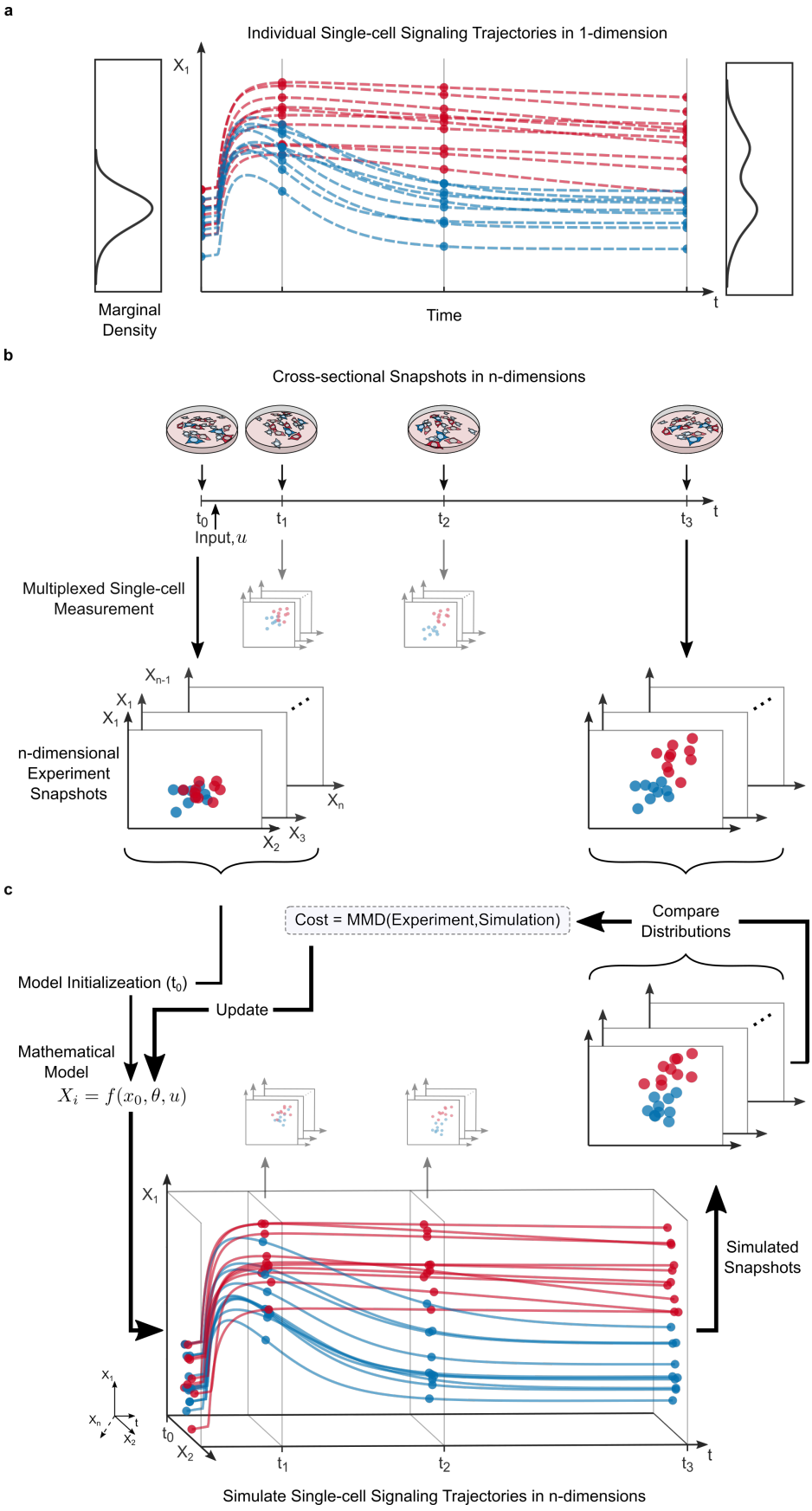

a
Individual Single-cell Signaling Trajectories in 1-dimension
X1
Marginal Density
Time
t
b
Cross-sectional Snapshots in n-dimensions
t0
t1
t2
t3
t
Input, u
Multiplexed Single-cell Measurement
Xn-1
X1
X1
n-dimensional Experiment Snapshots
Xn
X2
X3
c
Cost = MMD(Experiment,Simulation)
Compare Distributions
Model Initializeation (t0)
Update
Mathematical Model
$X_i = f(x_0, \theta, u)$
Simulated Snapshots
X1
t0
X2
t1
t2
t3
t
Simulate Single-cell Signaling Trajectories in n-dimensions

Figure 3.1: Motivation and workflow of single-cell ODE approach. (**a**) Motivation. One-dimensional (X_1) time-lapse measurements (dots) cannot explain the cause of a bimodal signaling response. Red and blue represent two subpopulations. (**b**) Experimental workflow. Multiplexed cross-sectional experimental snapshots characterize a dynamical process of single cells in multiple dimensions, but cannot measure response. Bi-axial plots represent two-dimensional projections of n-dimensional distribution snapshots. (**c**) Computational workflow. Steady-state measurements at t_0 are used to initialize a set of ODE model instances, one for each cell. Differences between simulated and measured snapshots of n-dimensional cell state distributions are compared to optimize model parameters. (Time points t_1 and t_2 are also compared, but shown smaller for simplicity).

To characterize continuous multiplexed single-cell signaling trajectories and study the origins of cell-to-cell differences in signaling response, we developed a combined experimental and computational framework, called single-cell ordinary differential equation modeling or SCODEM (Figure 3.1b-c). Unlike parametric methods that use an ODE system to describe the trajectories of a distribution, SCODEM uses an ODE system to describe the trajectory of an individual cell. The simulations of many cells, then, combine to represent the trajectory of an empirical (rather than parametric) distribution, which can be compared to experimental data using non-parametric statistical methods. To quantify the distance between experimental and simulated snapshots and inform model fitness, we used maximum mean discrepancy (MMD) (Gretton et al., 2012a), a distribution-free two-sample test for multivariate distributions. The use of distribution-free statistics for model fitting has major advantages over previous (parametric) approaches to model cellular variation in signaling. First, it decouples the number of free model parameters from the number of cells or subpopulations. This makes any single-cell modeling problem similar in parameter number to a corresponding classical population model. The second advantage is use of information in the complex multi-dimensional structure of multiplexed single-cell data in a manner that does not depend on statistical assumptions or parameters. Thus, without assuming the shapes or numbers of cell populations, MMD can successfully discriminate between differences in distributions otherwise missed by parametric statistical methods (Figure 3.2). Finally, SCODEM is relatively easy to implement: single-cell models are instantiated using single-cell measurements, and simulation of many cells fits within ensemble modeling frameworks.

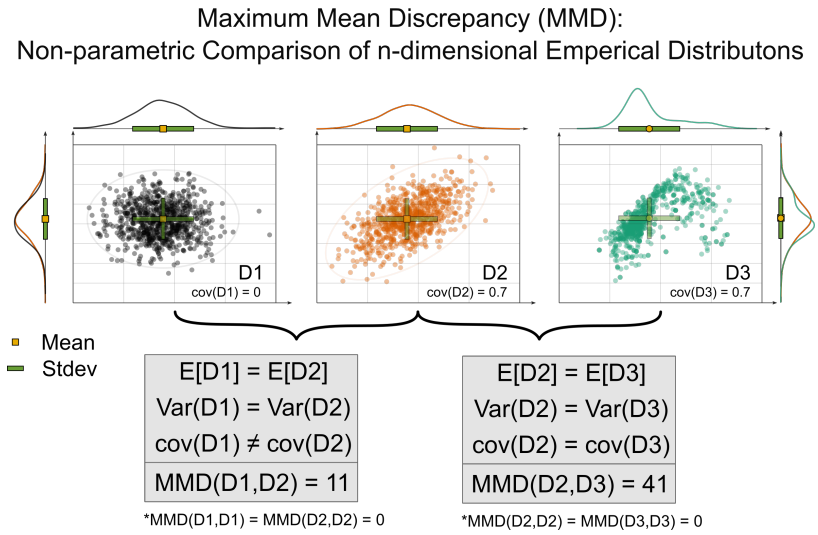


Figure 3.2: Advantage of non-parametric methods to compare distributions. Distributions D1 and D2 have equal mean and variance, but unequal covariance. D2 and D3 have equal mean, variance and covariance, but different higher-order structures. Methods based on these parameters fail to differentiate distributions D1, D2 and D3. MMD successfully differentiates D1, D2 and D3. Abbreviations: Var., variance; cov., covariance.

3.3.2 Multiplexed single-cell trajectories of the MAPK/ERK pathway signaling

Single-cell variation in signaling can be described deterministically.

MAPK/ERK pathway signaling is deregulated in many cancers, and pathway components are common drug targets (Samatar and Poulikakos, 2014). The success rates of pathway targeted therapies have been limited, however, at least in part due to variation in tumor cell responses (Wellbrock and Arozarena, 2016; Caunt et al., 2015). To study cell-to-cell differences in MAPK/ERK signaling responses, we used mass cytometry to measure multi-dimensional single-cell snapshots of cell states, including total and active MEK, ERK, and p90RSK, to characterize signaling at six time points during a one hour time course of HEK293T cells stimulated with EGF. Then, we used a subset of these data to construct an SCODEM model of the pathway. Although the true pathway involves many reactions, we condensed it to a model structure that minimized the number of parameters that cannot be obtained from experimental data, yet should have remained representative. The model includes active RAF and both the active and inactive states of MEK and ERK to represent the core pathway. The model also includes active and inactive p90RSK, which is a

downstream target of ERK (Figure 3.3a). For model fitting, we used the six-dimensional cell-state snapshot measurements of total and active MEK, ERK and p90RSK of approximately 500 cells subsampled from each time point. To qualitatively assess model fit, we simulated an independent subset of single cells and compared snapshots of the continuous trajectories to the experimental snapshot measurements. The strong agreement between both one-dimensional population statistics and distribution shapes (Figure 3.3b-c) and high-dimensional distribution structure, visualized by combinatorial two-dimensional projections for each time point (Figure 3.3d), confirm the model is able to represent the system for the conditions measured. This result shows that deterministic simulation of signaling in single-cells, based on a mechanistic model, is sufficient to reproduce the variation in signaling observed in multiplexed single-cell measurements.

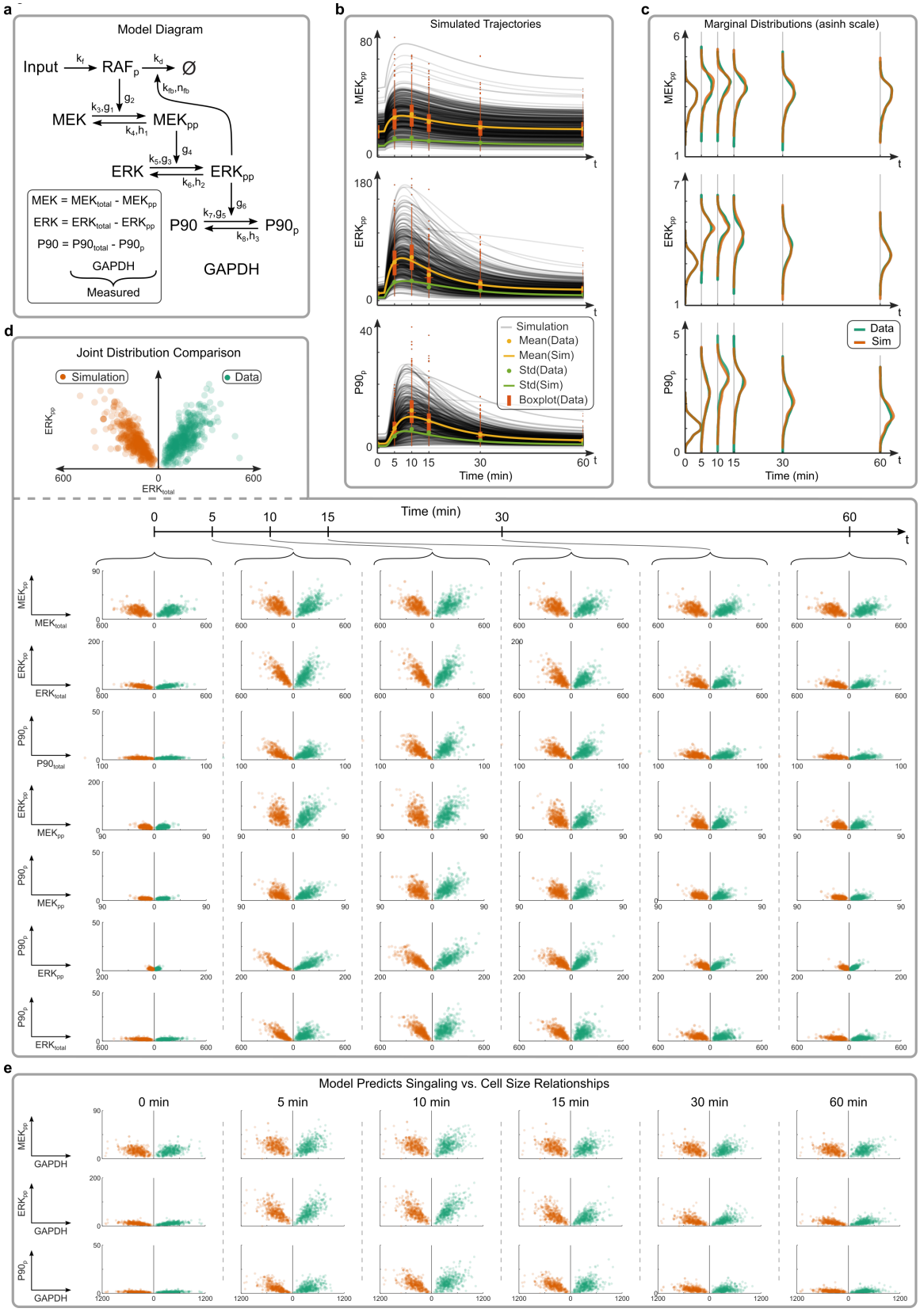

a
Model Diagram
Input
RAF_p
Ø
MEK
MEK_pp
ERK
ERK_pp
P90
P90_p
MEK = MEK_total - MEK_pp
ERK = ERK_total - ERK_pp
P90 = P90_total - P90_p
GAPDH
Measured
GAPDH
b
Simulated Trajectories
MEK_pp
ERK_pp
P90_p
Time (min)
Simulation
Mean(Data)
Mean(Sim)
Std(Data)
Std(Sim)
Boxplot(Data)
c
Marginal Distributions (asinh scale)
Data
Sim
d
Joint Distribution Comparison
Simulation
Data
ERK_pp
ERK_total
Time (min)
e
Model Predicts Singaling vs. Cell Size Relationships
0 min
5 min
10 min
15 min
30 min
60 min
GAPDH

Figure 3.3: Single-cell model of the RAF-MEK-ERK-p90RSK pathway. (**a**) Diagram of the pathway model annotated with kinetic parameters, which are to be inferred from measurements of total and active forms of MEK, ERK and p90RSK. (**b**) Simulated single-cell trajectories for active proteins compared to measured population statistics. Yellow and green circles (or lines) are mean and standard deviation of the data (or the simulations) at each presented time point. Orange boxplots are middle 50 (bar) and 90 (line) percent of data with a horizontal line at median, while events outside 90% range are dots. (**c**) Smoothed marginal (one-dimensional) distributions of data (green) and simulation (orange). (**d**) (Top left corner) Example plot comparing two-dimensional projections of multi-dimensional data (green) and simulated results (orange) at a given time point. Results along the X-axis are mirrored about 0, so that distance from origin always positive in the x-direction. Symmetry about the y-axis is a visual measure of good model fit. (Main panel) Summary of model fit to data in multidimensional space: Columns represent time points and rows exhibit different 2-dimensional projections of cell state. (**e**) Model validation. Model simulations reproduce the relationships between signaling proteins and GAPDH, which were not used in fitting. Relationships shown as in (d). Abbreviations: Sim., Simulation; Std, standard deviation. Model units: scaled concentrations (described in Appendix B).

Single-cell trajectories predict an additional cell-state dimension.

To validate our model, we predicted the multi-dimensional relationship between simulated signaling proteins and a measured cell-state variable not used in model fitting. Namely, we measured GAPDH abundance as a reference of cell volume (Rapsomaniki et al., 2018) (Figure B.1) but did not include this information in our model calibration, rather assuming each cell to have the same size (see Appendix B). GAPDH does not change during the time-period of the experiment (Figure B.2) and should not directly affect the signaling reactions. It can therefore be represented in the model as a constant for each cell during simulation (Figure 3.3a). To test the predictive ability of our model, we simulated continuous signaling trajectories in cells (starting at steady state) and compared snapshots of the relationships between GAPDH and signaling proteins with those from independent snapshot experiments (Figure 3.3e). Because GAPDH was not used in modeling fitting, these joint relationships between signaling proteins and GAPDH in simulated trajectories are true predictions of the model. The clear agreement between model predictions and experimental snapshots (Figure 3.3e) lends strong support to the inferred single-cell trajectories and their subsequent use to quantify and analyze cellular responses (Figure 3.4).

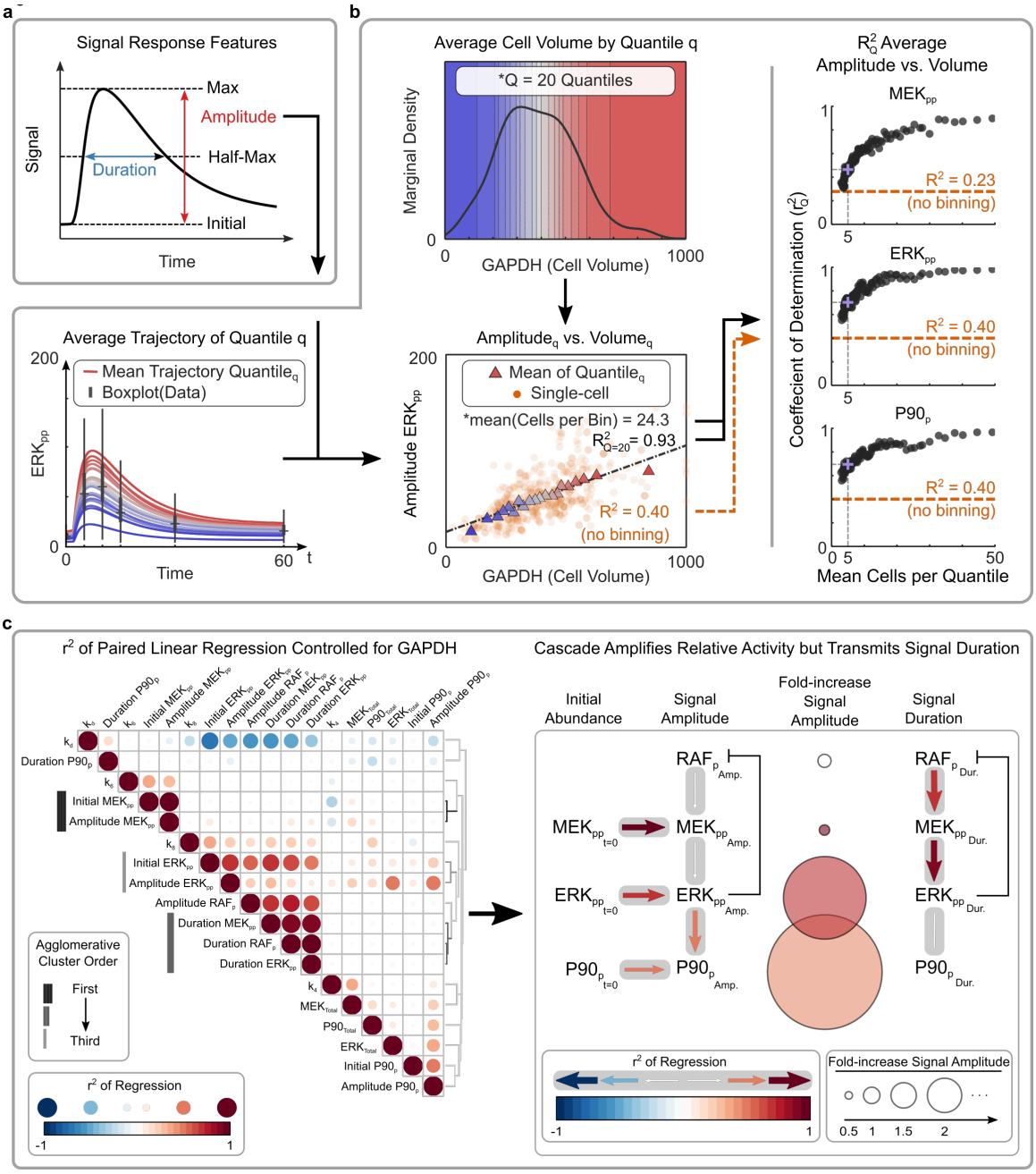

a
Signal Response Features
Max
Amplitude
Half-Max
Duration
Initial
Signal
Time
Average Trajectory of Quantile q
Mean Trajectory Quantile$_q$
Boxplot(Data)
ERK$_{pp}$
Time
b
Average Cell Volume by Quantile q
*Q = 20 Quantiles
Marginal Density
GAPDH (Cell Volume)
Amplitude$_q$ vs. Volume$_q$
Mean of Quantile$_q$
Single-cell
*mean(Cells per Bin) = 24.3
$R^2_{Q=20}$= 0.93
R^2 = 0.40
(no binning)
Amplitude ERK$_{pp}$
GAPDH (Cell Volume)
R^2_Q Average Amplitude vs. Volume
MEK$_{pp}$
R^2 = 0.23
(no binning)
ERK$_{pp}$
R^2 = 0.40
(no binning)
P90$_p$
R^2 = 0.40
(no binning)
Coeffecient of Determination (r^2_Q)
Mean Cells per Quantile
c
r^2 of Paired Linear Regression Controlled for GAPDH
Agglomerative Cluster Order
First
Third
r^2 of Regression
Cascade Amplifies Relative Activity but Transmits Signal Duration
Initial Abundance
Signal Amplitude
Fold-increase Signal Amplitude
Signal Duration
RAF$_p$ Amp.
MEK$_{pp}$ t=0
MEK$_{pp}$ Amp.
ERK$_{pp}$ t=0
ERK$_{pp}$ Amp.
P90$_p$ t=0
P90$_p$ Amp.
RAF$_p$ Dur.
MEK$_{pp}$ Dur.
ERK$_{pp}$ Dur.
P90$_p$ Dur.
r^2 of Regression
Fold-increase Signal Amplitude

Figure 3.4: Sources of single-cell variation in signaling response. (**a**) Definition of signal response features for each single-cell trajectory. Amplitude: maximum absolute increase in signal. Duration: length of time-period where signal is greater than half-maximal. (**b**) Averaging cells by quantile of GAPDH expression shows the strong linear relationship between cell volume and signal amplitude (example using Q = 20 quantiles shown). (Upper Middle) Smoothed histogram of GAPDH (black line) divided into 20 quantiles by GAPDH expression. The colored panels represent each quantile from low (blue) to high (red) GAPDH expression. Cells in each quantile bin are averaged for the analysis. (Lower Left) Average trajectory of cells in GAPDH quantile q colored by quantile. (Lower Middle) Regression of average amplitude versus average GAPDH (triangles) colored by quantile q of GAPDH. Orange dots are individual-cell observations. (Right) A summary of the r^2_Q of regression for each signaling protein as a function of median cells per quantile (based on number of quantiles Q) of GAPDH. The violet cross intersected by gray dashed lines represents R^2_Q five cells per quantile. Horizontal orange dashed line is R^2 of regression using single-cell values (i.e. without averaging). (**c**) ERK cascade amplifies relative activity and transmits signal duration. (Left) r^2 of partial linear regression between all pairs of single-cell features, controlling for GAPHD by partial least squares. (Right) Graphical summary of key information using MAPK/ERK pathway structure. Amplitude depends on initial state, but is not transmitted through the ERK cascade. Signal duration is transmitted through cascade. Circles represent fold-increase in signal and are colored by R-squared between initial value and signal amplitude of protein at corresponding level of the pathway. Abbreviations: -p and -pp as in Figure 3.3a; Dur., Duration; Amp., Amplitude. Model units: scaled concentrations.

3.3.3 Sources of variation in signaling response of the MAPK/ERK pathway

Variation in ERK pathway signaling has been attributed to extrinsic (unmeasured) variation upstream of MEK (Filippi et al., 2016). The lack of continuous multiplexed measurements, however, has left the source of variation in signaling an open question. To determine sources of variation in ERK pathway signaling, we used the multiplexed single-cell signaling trajectories provided by our model. We used peak amplitude and duration (Figure 3.4a) as metrics of signaling response for each pathway component RAF, MEK, ERK and p90RSK. We then added the eight resulting signal response features (two features times four signaling proteins) to the state space of each cell at the steady state to generate a 19-dimension distribution of cell state and signaling responses. Using regression, the combined cell-state-cell-response distribution was able to explain major sources of variation in the response metrics of the ERK pathway.

Cell volume explains a large portion of the single-cell variation in peak amplitude.

We hypothesized that cell volume could confound results based on abundance measurements. To test this, we quantified the relationship between signal amplitude and cell volume using paired linear regression of each signaling protein versus GAPDH. The resulting R^2 values show that GAPDH explains approximately 40% ($R^2 = 0.40$) of the variation in ERK and p90RSK amplitude, and nearly one quarter of the variation in MEK amplitude ($R^2 = 0.23$). To determine the sensitivity of these relationships to measurement error (as GAPDH is only an approximate measure of cell volume, for example), we grouped and averaged cells with similar GAPDH abundance (by quantile) and repeated the regression. We found that the coefficient of determination as a function of the cells per quantile (R^2_Q, where Q equals the number of quantiles) rapidly increased when averaging even few cells, which illustrated the strong linear relationship between cell volume and amplitude (Figure 3.4b). For example, averaging groups of approximately five cells by GAPDH abundance (using $Q = 145$ quantiles) increased the R^2_Q between GAPDH and the signaling amplitude of MEK, ERK and p90RSK to 0.46, 0.71 and 0.69, respectively (Figure 3.4b). Because each cell in our model was assumed to have the same size, the linear relationship between cell volume and amplitude could be directly explained by concentration, which equals abundance/volume. Thus, cell volume does confound abundance measurements and concentration-based measurements would greatly reduce the observed single-cell variation in signaling amplitude compared to measurements of protein abundance.

Initial phosphorylation state is the best predictor of signal amplitude.

The variation in phospho-protein abundance at peak signaling times is a classical example of cell-to-cell variation in signaling. To study the relationships between signaling amplitude and pathway components, we calculated the coefficient of determination (r^2) for each pair of cell-state and response features, controlling for GAPDH (Figure 3.4c). For each signaling protein, MEK, ERK and p90RSK, the initial activity level strongly predicted the signaling amplitude ($r^2 > 0.5$). Thus, the fold-increase in signal, defined as peak amplitude divided by initial activity, was much less variable than peak signal amplitude across the cell population.

For ERK, this agrees with previous live-cell studies (Cohen-Saidon et al., 2009). Indeed, we found initial activity was the best of all possible predictors of signaling amplitude for MEK and ERK (Figure 3.4c). These results show that, starting from a steady state where signaling is constant, the ERK signaling cascade is tuned such that each node amplifies EGF signal relative to steady-state, independent of the upstream signaling amplitude, and accounting for the normal steady-state level greatly reduces variation in signaling amplitude across a population.

MAPK/ERK signaling cascade transmits signal duration, not amplitude.

Signal strength and duration are commonly presented as being transmitted by the ERK pathway (Shaul and Seger, 2007). Somewhat contrary to this idea, however, we found that the amplitudes of MEK and ERK are not strongly correlated to the amplitude of their upstream activators RAF ($r^2 = 0.08$) and MEK ($r^2 = 0.08$), respectively (Figure 3.4c). Thus, signaling amplitude is not reliably transmitted by the MAPK/ERK cascade. The amplitude of downstream p90RSK, however, does depend on the amplitude of ERK ($r^2 = 0.66$) slightly more than on any other cell feature. In contrast to amplitude, the signal duration of RAF, MEK and ERK are all tightly coupled ($r^2 > 0.89$; (Figure 3.4c). The duration of p90RSK signaling, however, is not obviously predicted by any one cell feature. The high-fidelity transmission of signal duration through the core RAF/MEK/ERK signaling unit, combined with the low transmission of information on peak amplitude in favor of constant fold-change signal amplification, shows that the MAPK/ERK signaling cascade operates in a way to robustly transmit signal duration, rather than signal amplitude.

3.3.4 The model predicts cell states insensitive to drug treatment.

Because variation in cancer cell state, exclusive of genetic variation, can lead to corresponding variation in drug response and treatment failure (Shaffer et al., 2017), we tested the ability of our model to predict variation in signaling during acute inhibition. We simulated an experiment treating cells for 30 minutes prior to EGF stimulation with CI-1040 (PD184352), a once promising small-molecule MEK inhibitor that however failed to show sufficient anti-tumor activity in a phase II clinical trial (Chang-Yew Leow et al., 2013).

Specifically, CI-1040 works by non-competitively (i.e., independent of kinase-substrate binding) blocking the ability of active MEK to phosphorylate ERK (Allen et al., 2003). We used the biochemical interpretation of non-competitive inhibition (Sebolt-Leopold et al., 1999) to simulate EGF signaling under a range of CI-1040 concentrations (Appendix B; Figure B.3). Using fold-increase in phospho-ERK to quantify cell activation, the model revealed cells that remained responsive to EGF under non-saturating concentrations of the MEK inhibitor (Figure B.3). To test this model prediction, we performed the corresponding experimental inhibition/stimulation time-course. Although the time-scale of activation was slower than that predicted by our model, a subset of CI-1040 treated cells showed increased phospho-ERK after EGF stimulation as predicted.

Inhibition has secondary kinetic effects.

We hypothesized the difference in dynamics between simulations and experiments was due to secondary or indirect effects of the inhibitor. Using systematic model analysis and optimization, we identified a slightly expanded kinetic model of the inhibitor that represents the inhibition data well (Figure 3.5; Figure B.4). In this expansion, the inhibitor slows down not only the activation rate of ERK (k_5), as classically expected, but also the inactivation rate of ERK (k_6) and both the activation (k_3) and inactivation (k_4) rates of MEK (Figure 3.5a). Subsequently, we found that CI-1040 can indeed inhibit MEK activation by RAF (Kramer et al., 2004). This reconciliation result illustrates the ability of SCODEM to identify unintuitive secondary effects of kinetic modulators in the context of a complex cellular system.

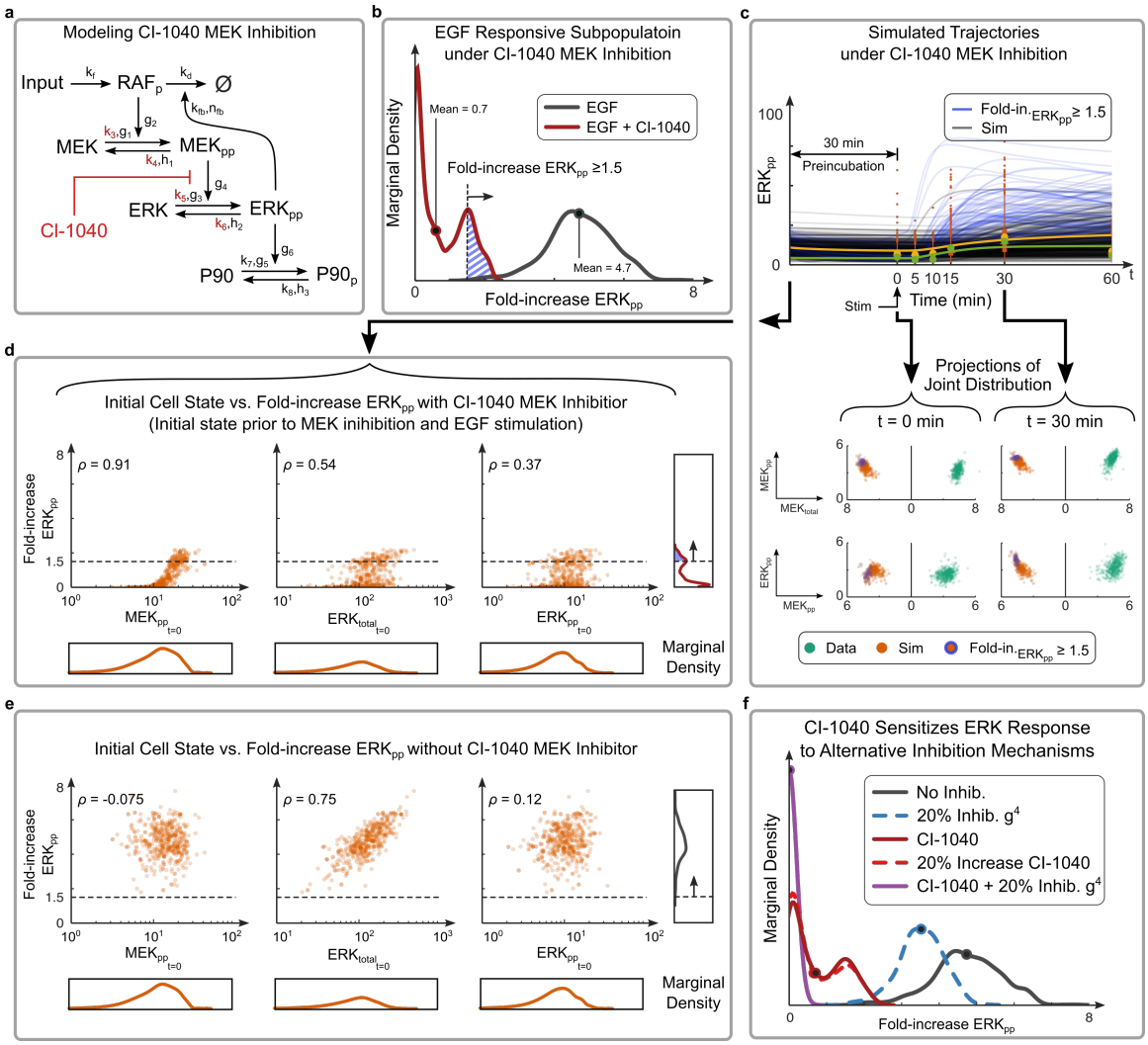


Figure 3.5: Modeling MEK inhibition reveals characteristics of insensitive cells. (**a**) Model diagram with parameters (red) modified by MEK inhibitor Cl-1040. (**b**) Distributions of fold-increases of ERKpp in EGF treated cells both with (red) and without (gray) 30 min pre-incubation of CI-1040 MEK inhibitor. Fold-increase in signal amplitude $\geq$ 1.5 for Cl-1040 treated cells (blue hatch) and both population means (circles) shown. (**c**) (Top panel) Simulated trajectories of ERKpp with 30-min pre-incubation of Cl-1040. Trajectories with relative response $\geq$ 1.5 (chosen as greater than the minimal response of stimulated cells not treated with inhibitor) in blue, all others gray. Summary statistics as in Figure 3.3b. (Bottom panel) Example snapshots of data and simulation at 0 and 30 min stimulation time-points, visualized as in Figure 3.3d (full 60-minute time course in Figure B.4). Orange circles with blue outline show cells with ERKpp response $\geq$ 1.5. (**d,e**) Initial cell state (x-axis) vs. relative signaling amplitude Erkpp with (d) and without (e) CI-1040 MEK inhibition (y-axis). Projection of marginal density of each axis and Spearman's rho shown. Dashed line corresponds to fold-increase Erkpp$\geq$ 1.5, as in (b). (**f**) Distributions of fold-increase in active ERK (as in b) for simulated treatment conditions. Circles represent population mean. Kinetic parameter g^4 shown in (a). Abbreviations: -p and -pp as in (a); Sim., Simulation. Model units: scaled concentrations.

Drug treatment sensitizes ERK response to initial MEK activity levels.

Contrary to snapshot measurements, the continuous multiplexed trajectories from our model allowed us to identify not only a subset of cells insensitive to drug treatment (fold-increase of active ERK greater than 1.5; Figure 3.5b-c), but associated initial cell states that we used to predict single-cell drug response. To do this, we compared the fold-increase in ERK signal to measurable cell features at steady state (before inhibitor treatment). The model predicts that cells insensitive to CI-1040 inhibitor have high initial levels of active MEK (MEKpp; Figure 3.5d). In cells not treated with inhibitor, however, initial MEK activity did not predict ERK response (Figure 3.5e). Thus, our model predicts that use of CI-1040 sensitizes the ERK response to MEK activity. Mechanistically, this is driven by the ultrasensitive (Huang and Ferrell, 1996) (highly nonlinear) response of ERK to active MEK (response curve in Figure 3.5d; driven by parameter g^4 in Figure 3.5a; Appendix B), which dominates the ERK activation reaction when CI-1040 treatment otherwise reduces the forward reaction rate constant (parameter k_5 is decreased; Appendix B). Importantly, the ultrasensitivity of ERK response to MEK activity induced by CI-1040 may be leveraged to further inhibit ERK signaling via a complementary inhibitory mechanism. For example, a competitive MEK inhibitor or MEK/ERK scaffold inhibitor would be expected to decrease the ultrasensitivity in the relationship between MEK and ERK (captured in parameter g_4). While neither a 20% decrease in g_4 nor a further 20% increase in CI-1040 would meaningfully increase inhibition of ERK activation, combination of a 20% decrease in g_4 with the current CI-1040 treatment would effectively abolish ERK response (Figure 3.5f). These results show how SCODEM can be used to identify cell states less sensitive to drug treatment, and, with subsequent analysis, how drugs targeting the same molecule, but different mechanisms, may be used in a cooperative fashion.

3.3.5 Effects of ERK overexpression on MAPK/ERK signaling.

Many cancers are caused by mutations (Blume-Jensen and Hunter, 2001) and/or environmental changes (Abu-Remaileh et al., 2015; Jeong et al., 2014) that greatly increase the range of protein expression in a cell population and thereby alter signaling behavior (Lun

et al., 2017). As overexpression of a known signaling protein should not change the structure (kinetic rate functions) of the signaling system, we hypothesized that related signaling changes should be a function of expression differences alone. We used SCODEM to test this. Our model, however, was constructed primarily with power-functions to represent reaction kinetics. Power-functions are local approximations of the "true" reaction functions and well suited for normal operating ranges of protein expression, but may fail to represent large ranges of expression as they do not saturate and reaction functions clearly cannot increase without limit. (Savageau, 1976; Voit, 2013) Therefore we added saturating functions to the model to limit the range of reaction rate increase (Figure 3.6a).

With this alteration, we modeled an experimental protein overexpression system (Lun et al., 2017), based on transient transfection of ERK2-GFP, which randomly introduces additional ERK to cells and increased the range of ERK expression by two orders of magnitude (Figure 3.6b). We then measured the signaling dynamics in response to EGF stimulation. To account for our assumption that high values of ERK could saturate reactions, we used these overexpression data to fit the added parameters. As a validation, we independently simulated the combination of overexpression and stimulation experiments. Initiating the model with a sample of cells at steady state taken from non-overexpression experiments, we simulated transfection by random addition of ERK to each cell according to a representative distribution (Figure 3.6c; Appendix B). Next, we simulated EGF signaling. As before, we qualitatively compared snapshots of the simulated and experimental six-dimensional signaling state distributions at each time point using combinatorial projections (Figure 3.6d). The striking resemblance of the complex distributions in simulated and experimental snapshots supported subsequent analysis and illustrates that even highly heterogeneous cell samples with complex multi-dimensional dynamics can be represented deterministically, given initial cell states.

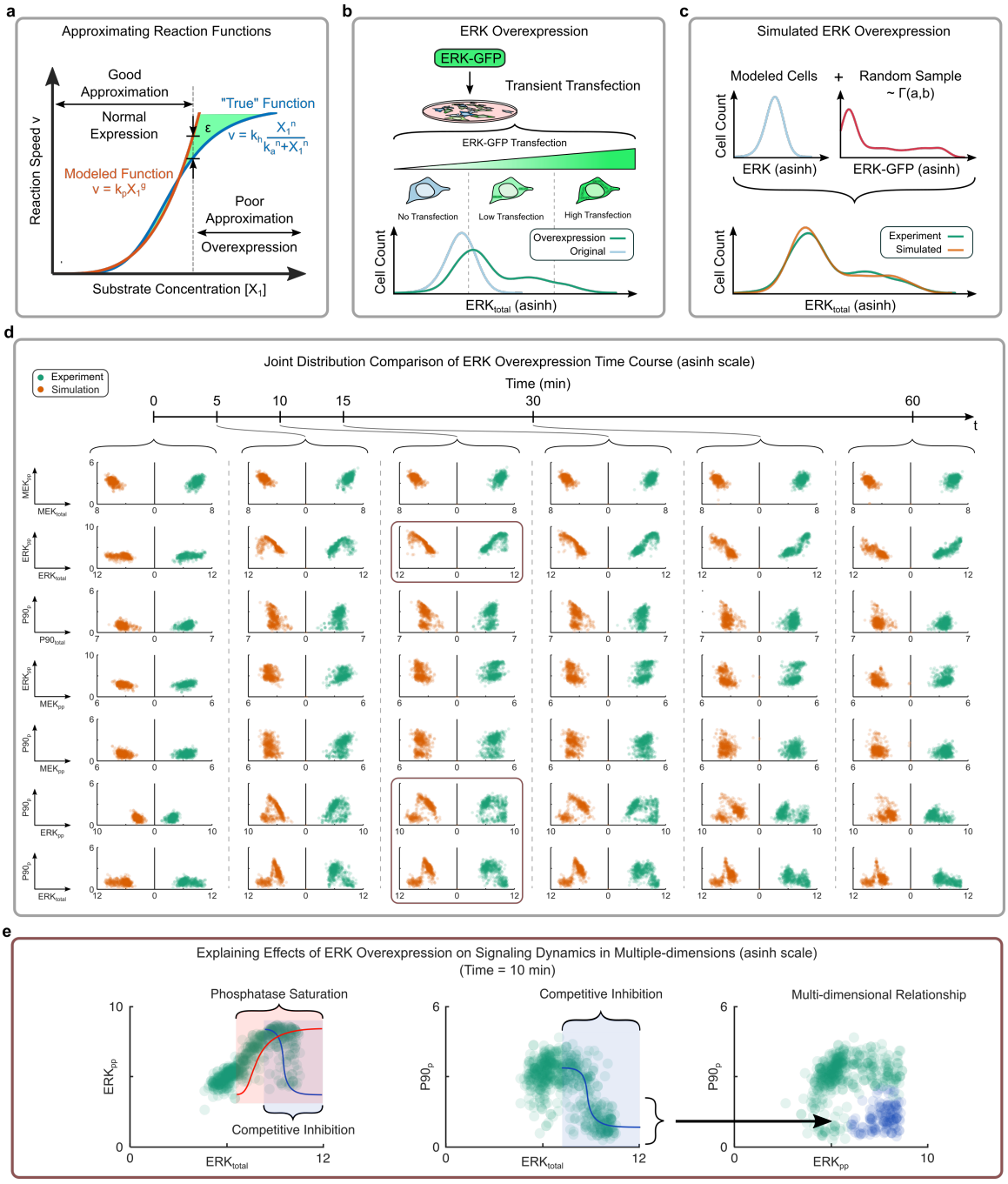

a
Approximating Reaction Functions
Good Approximation
Normal Expression
"True" Function
ε
Modeled Function
Poor Approximation
Overexpression
Reaction Speed v
Substrate Concentration [X1]
b
ERK Overexpression
ERK-GFP
Transient Transfection
ERK-GFP Transfection
No Transfection
Low Transfection
High Transfection
Overexpression
Original
Cell Count
c
Simulated ERK Overexpression
Modeled Cells
Random Sample
~ Γ(a,b)
ERK (asinh)
ERK-GFP (asinh)
Experiment
Simulated
d
Joint Distribution Comparison of ERK Overexpression Time Course (asinh scale)
Experiment
Simulation
Time (min)
e
Explaining Effects of ERK Overexpression on Signaling Dynamics in Multiple-dimensions (asinh scale)
(Time = 10 min)
Phosphatase Saturation
Competitive Inhibition
Multi-dimensional Relationship

Figure 3.6: Kinetic effects of ERK overexpression. (**a**) A non-saturating power-law function (red) is often a good approximation of a saturating kinetic function (blue) throughout a "normal" range of substrate concentration, yet fails when the substrate concentration is increased beyond the normal range. Green space represents error (ϵ) between functions. (**b**) Experimental transient transfection of ERK2-GFP results in a wider distribution of total ERK (green) compared to control (blue). (**c**) Simulated random addition of ERK (sampled from a censored gamma distribution (red)) to control cells (blue) in the model (orange) closely resembles the ERK distribution of experimental overexpression (green). (**d**) Comparison of simulated (orange) and measured (green) time-course of EGF stimulation under ERK overexpression (format as in Figure 3.3d). See Figure B.5 for single-cell trajectories and marginal distributions as in Figure 3.3b-c. Red boxes are distributions shown in Panel (e). (**e**) Explanation of specific distributions in red boxes from Panel (d). (Left and Center) Illustration of kinetic mechanisms that generate distribution shapes. (Right) illustrates the connection between the phoshpo-p90RSK-low subpopulation across different projections of the state-space. Abbreviations: Sim., simulation; Fold-in., fold-increase; Inhib., inhibition. Model units: scaled concentrations.

We analyzed our model to determine the consequences of ERK overexpression on signaling and found that overexpression affects not just ERK activation, but the reaction kinetics of the entire MAPK/ERK pathway. The range of ERK overexpression where each effect is dominant, however, varied (Figure 3.6e). Specifically, beginning at moderate levels of overexpression, the reaction deactivating ERK (i.e. the ERK phosphatase) begins to saturate and leads to greatly increased levels of active ERK. Moderate to high levels of overexpression reduced the strength of the negative feedback from ERK to RAF, which slows the decay in input signal. At high levels of ERK overexpression, both the activating and inactivating reactions of MEK and ERK were slowed, an effect consistent with competitive inhibition, and also leads to longer signal duration of active ERK. Surprisingly, high levels of ERK also completely inhibited p90RSK activation in cells with slowed ERK activation. Because p90RSK activation requires dissociation from inactive ERK (Roux et al., 2003), however, competition (Levchenko et al., 2000) between active and inactive ERK for inactive p90RSK may explain the inhibitory effect of high ERK overexpression. As saturating functions were generally not needed to represent "normal" cells (as in Figure 3.3), our results show how protein overexpression can have complex kinetic effects, and these effects cannot necessarily be observed in "normal" cells, even with single-cell measurements. Our results also show the ability of SCODEM to capture complex kinetics from appropriate measurements for use

in modeling, even when measurements are made in a mixed cell population.

3.4 Discussion

The trade-off in single-cell experiments between measuring either the dynamics of a few cell states or snapshots of many cell states has been a barrier to understanding how cell-to-cell variation in expression affects signaling dynamics. Here, we present a combined experimental and computational approach (SCODEM) to infer multiplexed signaling trajectories from snapshot data, thus bridging this gap. We applied this approach to EGF signaling in the MAPK/ERK pathway of HEK cells. The highlights are: (1) cell-to-cell variation in signaling can be described deterministically, given the initial state of each cell; (2) cell volume is a significant contributor to the variability observed in single-cell measurements; (3) signal duration is reliably transmitted through the RAF/MEK/ERK cascade, whereas signal amplitude is a fold-change response. More generically, we demonstrate that SCODEM enables us to study variability in cellular responses to drug treatments and in disease.

Use of SCODEM requires several considerations. The first is genetic variation within a sample. In this study, we assume a monoclonal cell population, which justifies identical biochemical parameters across cells. In cases where enzymatic activity, in addition to expression, drive differences in response, this assumption might fail. However, the parameters in the kinetic models used in SCODEM have a direct functional interpretation, which helps prevent unrealistic choices of numerical values. Furthermore, single-cell measurements of both protein expression and genotype may be used to inform enzymatic parameters in these cases. A second consideration is the direct reliance on single-cell measurements as model inputs, which can increase errors in estimates for low signal ranges and complicate normalization across channels (Appendix B). Third, our distribution-free approach to model fitting should also be applicable gene network models, although stochastic reaction kinetics may be necessary to simulate single-cell trajectories. Finally, the ease of applying SCODEM to larger networks is determined primarily by the ability to measure the initial (steady-state) values of the model's state variables (i.e., network components), as unmeasured state variables necessarily have values that must be either assumed or inferred (Appendix B).

Single-cell snapshot measurements reveal variation in cell states. Our question, however, is how variation in cell states relates to variation in cell responses. Here, we show for the MAPK/ERK pathway that: (1) the variation in signaling responses is much less than variation in the signaling components; and (2) the variation in unmeasured signaling components can be algebraically determined by appropriately analyzing experimental measurements with a deterministic model (Appendix B). These results suggest that the structures of both the signaling and expression systems cooperate to reduce variation of cell responses within a population. A clear example is the strong correlation between cell volume, protein abundance and signaling amplitude, which is reasonably explained by isometric scaling of signaling components as cells grow. Expressed differently, a large amount of variation in single-cell measurements is explainable and therefore coordinated, and signaling systems are structured in such a manner that cells can tolerate the remaining noise and mount appropriate responses in a robust manner. The result secondarily suggests caution before unexplained variation in cellular systems is attributed to randomness.

While our results demonstrate the well-coordinated maintenance of signaling response characteristics, they also show that this coordination can be impaired in diseases such as cancer, which are associated with mutations that alter the cellular control of protein expression. We demonstrate such system failure with an overexpression system that reveals how cells operating in abnormal regions of the expression space can have drastically altered signaling response characteristics. Critically in these cases, "normal" cells operate within their physiologically constrained state-space, whereas "mutated" cells enter new state-space regions where reaction functions and consequent cellular responses might qualitatively change behavior. This insight suggests that cellular systems must be studied both within and outside their normal operating conditions to enable reliable predictions of signaling responses in disease. Such analyses of signaling in deregulated systems are facilitated by tightly connected experimental and computational analyses, for instance, by combining a protein overexpression system with mass cytometry and SCODEM as done here.

Predicting how individual cells in a tumor will respond to treatment is substantially confounded by single-cell variation in state and environment. Additionally, observations

of primary tumor cells are generally limited to snapshot measurements. We show how an appropriately calibrated single-cell ODE model may be used to predict single-cell drug responses from single snapshot measurements, and how one inhibitor may subsequently sensitize cells to another inhibitor with the same target but a different mechanism of action. We demonstrate this capability only with a rather simple model and with measurements under culture conditions, but our study nonetheless represents proof of principle that single-cell ODE modeling can deepen our understanding of differential cellular responses.

3.5 Methods

Quantifying ERK pathway signaling dynamics in response to EGF

Samples were fixed at {0,5,10,15,30,60} minute time points after stimulation with EGF and measured by mass cytometry (the 0-minute sample was not stimulated). Time points were chosen as appropriate to characterize EGF signaling dynamics in the pathway (Lun et al., 2017). Abundance of both total and active proteins were measured for the core pathway components MEK, ERK and p90RSK, as well as markers for cell volume (GAPDH), cell cycle (IdU, Cyclin B1, pHH3) and cell death (cleaved PARP), were measured using a validated panel of antibodies. The full antibody panel is in Table B.4.

Cell culture

HEK293T cells, obtained from ATCC, were cultured in DMEM (D5671, SIGMA), supplemented with 10% FBS, 2 mM L-glutamine, 100 U/ml penicillin, and 100 μg/ml streptomycin. For cell passaging or harvesting, cells were incubated with 1$\times$ TrypLE Express (Life Technologies) for 2 minutes at 37 °C. Purity and sterility of the cell line were certified by ATCC. Mycoplasma was not detected with the LookOut Mycoplasma PCR Detection Kit (Sigma-Aldrich).

Cloning

DNA sequences of the genes of interest were provided in entry clones by William Hahn and David Root (Yang et al., 2011) (via Addgene and NEXUS Personalized Health Technologies at ETH Zurich). Destination vectors, including pDEST pcDNA5 FRT TO-eGFP,

pDEST 5' Triple Flag pCDNA5 FRT TO and pDEST 3' Triple Flag pcDNA5 FRT TO, were kindly provided by Anne-Claude Gingras at the Lunenfeld-Tanenbaum Research Institute, Toronto, Canada (Couzens et al., 2013). Expression vectors encoding the FLAG- or GFP-tagged fusion proteins were generated via Gateway Cloning and sequenced before transfection.

Transfection and stimulation

HEK293T cells were seeded at a density of 0.7 million cells per well in 6-well plates. After 24 hours, cells were transfected with 2 μg plasmid and 4 μl of jetPRIME (PolyPlus) per well with the standard protocol provided by the manufacturer. At 18 hours after transfection, EGF (Peprotech) was added to a final concentration of 100 ng/ml. At 20 minutes before a given EGF stimulation time point, 5-Iodo-2-deoxyuridine (IdU) was added to the medium at the final concentration of 10 μM. At 2 minutes before a given EGF stimulation time point, medium was replaced by 1$\times$ TrypLE to induce cell detachment. At this time point, paraformaldehyde (PFA, from Electron Microscopy Sciences) was added to the cell suspension to a final percentage of 1.6%, and cells were incubated at room temperature for 10 minutes. If EGF stimulation was not necessary in the experiment, cells were directly harvested and crosslinked with PFA. Crosslinked cells were washed twice with cell staining media (CSM, PBS with 0.5% BSA, 0.02% NaN_3) and after centrifugation, ice-cold methanol was used to resuspend the cells, followed by a 10-minute permeabilization on ice or for long-term storage at -80°C.

MEK inhibition by CI-1040

The MEK1/2 inhibitor CI-1040 (Selleckchem) was pre-dissolved in DMSO at a concentration of 10mM and added to the cell culture plates at the final concentration of 0.5 μM for 30 minutes before EGF stimulation. Cell stimulation and harvesting were performed as described above.

Antibody conjugation

The MaxPAR antibody conjugation kit (Fluidigm) was used to generate isotope-labeled antibodies using the manufacturer's standard protocol. After conjugation, the antibody yield was determined based on absorbance of 280 nm. Candor PBS Antibody Stabilization solution (Candor Bioscience GmbH) was used to dilute antibodies for long-term storage at 4°C.

Barcoding and staining protocol

Formalin-crosslinked and methanol-permeabilized cells were washed three times with CSM and once with PBS. Cells were incubated in PBS containing barcoding reagents (^{102}Pd, ^{104}Pd, ^{105}Pd, ^{106}Pd, ^{108}Pd, ^{110}Pd, ^{113}In and ^{115}In) at a final concentration of 100 nM for 30 minutes at room temperature and then washed three times with CSM (Bodenmiller et al., 2012). Barcoded cells were then pooled and stained with the metal-conjugated antibody mix at room temperature for 1 hour. The antibody mix was removed by washing cells three times with CSM and once with PBS. For DNA staining, iridium-containing intercalator (Fluidigm) diluted in PBS with 1.6% PFA was incubated with the cells at 4°C overnight. On the day of the measurement, the intercalator solution was removed, and cells were washed with CSM, PBS, and ddH_2O. After the last washing step, cells were resuspended in ddH_2O and filtered through a 70-μm strainer.

Mass cytometry analysis

EQ Four Element Calibration Beads (Fluidigm) were added to cell suspensions in a 1:10 ratio (v/v). Samples were analyzed on a Helios mass cytometer (Fluidigm). The manufacturer's standard operation procedures were used for acquisition at a cell rate of approximately 500 cells per second. After the acquisition, all FCS files from the same barcoded sample were concatenated (Bodenmiller et al., 2012). Data were then normalized, and bead events were removed (Finck et al., 2013) before doublet removal and de-barcoding of cells into their corresponding wells using a doublet-filtering scheme and single-cell deconvolution algorithm (Zunder et al., 2015). Cytobank (http://www.cytobank.org/) was used for additional gating

on the DNA channels (^{191}Ir and ^{193}Ir) and ^{139}La/^{141}Pr to remove remaining doublets, debris and contaminating particulates. Data were then exported as .fcs files for subsequent analysis.

Spillover correction

Due to panel design, channel-to-channel spillover as a function of mass resolution (1 and +1 channels) and oxidation (+16 channels) was only possible in two cases:

1. From channel ^{143}Nd (total ERK) to ^{144}Nd (ppMEK) and ^{159}Tb. (GAPDH)
2. From channel ^{149}Sm (total p90RSK) to ^{150}Sm (total MEK).

Spillover between channels ^{149}Sm and ^{150}Sm was not significant for the measured count ranges and ignored. Due to events with very high counts in ^{144}Nd (from ERK overexpression), spillover from ^{144}Nd was compensated using an estimated 2.2% spillover correction to channels ^{144}Nd (+1) and ^{159}Tb (+16).

Data normalization and scaling for use in modeling

Experimental measurements were normalized for comparison across independent experiments and measurements. Before use in modeling, measurement channels were linearly scaled to satisfy biological constraints and facilitate direct physical interpretation of resulting model parameter values. Cells used in fitting and visualization of results were subsampled to reduce unnecessary computational cost. A full description of normalization, scaling and subsampling is presented in Appendix B.

ODE integration and solver speed-up

The ODEMEX CVode wrapper for Matlab (Vanlier et al., 2012) was used to compile MEX files in C++ using numerical integrators from the SUNDIALS CVode package (Lawrence Livermore National Laboratory, Livermore, CA).

Parameter optimization

Parameter optimization was performed using a combination of global and local search methods. Given a user defined starting region, systematic global search with local refinement was

performed using the `ESS` and `MULTISTART` algorithms in the MEIGO optimization toolbox (Egea et al., 2014). Further local refinement was based on the unconstrained local search algorithm `fminsearch` in the Matlab Optimization Toolbox (The MathWorks, Inc.).

Maximum mean discrepancy (MMD)

Maximum mean discrepancy (Gretton et al., 2012a) (MMD) is a distribution-free method of testing the similarity between samples from two multivariate distributions, and has been shown as both computationally efficient and effective when comparing empirical samples from multivariate distributions (such as those derived from multiplexed single-cell data).

MMD represents the similarity between two distributions (e.g., experiment and simulation snapshots) as the distance between the mean embeddings of distribution features in a reproducing kernel Hilbert Space (RKHS). Depending on the kernel k associated with the RKHS, infinitely many features may be used to compare the distributions (in contrast to, for example, comparison of only a few features such as the mean and variance).

Let $\mathcal{H}$ be the unit ball in a RKHS with associated kernel k. Given m samples from a distribution X and n samples from a distribution Y, then an empirical estimate of the MMD between X and Y is

$$\mathrm{MMD}_b[\mathcal{H}, X, Y] = \left[\frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n^2} k(y_i, y_j) \right]. \quad (1)$$

MMD as defined in equation (1) is a biased test statistic, but will still be small if $X = Y$ (the true distributions are equal) and large if the distributions are far apart. Bootstrapping can be used to generate a null distribution and determine the statistical significance of the difference between the distributions, but is computationally expensive if performed within an optimization routine. In this work, we are concerned with finding a model that generates distributions similar to experimental measurements, but are less concerned with the statistical significance of the model fit at any given step of the optimization algorithm (or comparison of two alternative models). Thus, in the SCODEM procedure presented here, it is sufficient to minimize the MMD between simulated and experimental distributions,

without additional bootstrapping to determine statistical significance. We note that boot strapping may be used to compare models after optimization, if desired.

Optimization objective (cost) function

Given discrete snapshot measurement at times $t \in T$, corresponding single-cell multiplexed snapshot experimental measurements D_t of v variables (e.g., proteins) in n cells ($\mathrm{D}_t \in \mathbb{R}^{n \times v}$), and a set of n continuous single-cell simulated trajectories $Y(t)$ of the v measurement variables, the objective function Cost is defined as:

$$\text{Cost} = \sum_{t \in T} \text{MMD}_b[\mathcal{H}, Y(t), D_t] \tag{2}$$

where kernel k associated with $\mathcal{H}$ (equation 1) is the Gaussian kernel:

$$k(x, x') = e^{-\frac{||x-x'||^2}{2\sigma^2}}. \tag{3}$$

The parameter σ scales the width of the kernel and was chosen as the median distance between points in the aggregate sample, which is the classical median distance heuristic for selection of kernel bandwidth. For our calculations, the individual observations $Y(t)$ and D_t were transformed using a hyperbolic arcsin function (asinh). The MMD computation was implemented in Matlab (The Mathworks) using code available at `http://www.gatsby.ucl.ac.uk/~gretton/mmd/mmd.htm`.

Mathematical model of EGF signaling in the MAPK/ERK pathway

The ODE model structure included an input, RAF, MEK, ERK and p90RSK, as well as the known negative feedback loop from ERK to active RAF (active ERK leads to removal of active RAF). The input was modeled as a single bolus corresponding to addition of EGF. Reaction steps in the MEK/ERK/p90RSK cascade were simplified to include molecular states that were experimentally observable. For example, ERK is activated when it is doubly phosphorylated on threonine 202 and tyrosine 204 (pT202/pY204) by active MEK in a two-step process. However, only the active pT202/pY204 form of ERK was measured,

and the transition from inactive to active ERK was modeled as a single step. The full model structure, equations and description are given in the Appendix B, as are the description of CI-1040 in the model and the details of model expansion to represent ERK overexpression.

All algorithms were implemented in Matlab Release 2015b (The MathWorks, Inc.).

Model fitting

The 17 free parameters of the model were trained on a representative subsample of approximately 500 single-cells from snapshot mass cytometry measurements during a 60-minute time-course of EGF stimulation. Specifically, the six-dimensional cell-state distribution of both total and active forms of MEK, ERK and p90RSK were used to inform model parameters.

Hyperbolic arcsin transformation

Where noted, data in some figures were transformed using a hyperbolic arcsin (asinh) transformation, which displays values between 0 and 1 on an approximately linear scale, and values greater than 1 on a log-like scale. This was done to better visualize highly dispersed distributions.

Signaling metrics

The amplitude and duration of signaling protein X_i in cell j during a time course T were defined as:

$$\text{Amplitude}(X_{i,j}) = \max_{t \in T}(X_{i,j,t}) - X_{i,j,t_0} \tag{4}$$

$$\text{Duration}(X_{i,j}) = \sum_{\{t \in T | Xi,j,t \geq \text{Amplitude}(X_{i,j})/2\}} X_{i,j,t} \cdot dt \tag{5}$$

In words, amplitude was the maximal increase in active protein and duration was the time period active protein was above the half-maximal level. Fold-increase in active protein (signal) was calculated as the amplitude normalized by steady state (the initial state where

$t = t_0$):

$$\text{Fold-increase}(X_{i,j}) = \frac{\text{Amplitude}(X_{i,j})}{X_{i,j,t_0}}. \tag{6}$$

CHAPTER IV

MECHANISTIC MODEL RECONCILES SIGNALING DYNAMICS ACROSS AN EPITHELIAL MESENCHYMAL TRANSITION[1]

4.1 Abstract

Intracellular signaling pathways are at the core of cellular information processing. The state of these pathways and their inputs determines signaling dynamics and drives cell function. Heterogeneous tissues, which are commonly encountered in cancer, comprise many combinations of cell states and microenvironments that can lead to variations in cell responses to treatments. Such context-dependent differences in signaling have traditionally been explained by network rewiring. However, biochemical rationale suggests that rewiring of signaling reaction networks should not be necessary. Here we address this conundrum with an *in vitro* model of an epithelial mesenchymal transition (EMT), a biological program implicated in increased tumor invasiveness, heterogeneity and drug resistance. We use mass cytometry to measure EGF signaling dynamics in the ERK/AKT signaling pathways before and after EMT and apply standard network inference methods, which clearly suggest EMT-dependent network rewiring. However, a more realistic modeling approach that adequately accounts for single-cell variation demonstrates that a single reaction-based pathway model with constant structure and near-constant parameters is sufficient to represent differences in EGF signaling across EMT. This result suggests that rewiring of the signaling network is not needed and that a unifying reaction-based model may be used to represent signaling in heterogeneous environments such as cancer.

4.2 Introduction

Intracellular signaling networks are biochemical systems that integrate spatio-temporal information regarding the intra- and extracellular state of a cell into functional programs

[1]This chapter is adapted from a manuscript and has the following author list: Wade, JD, Lun XK, Zivanovic N, Bodenmiller B, Voit EO.

that drive cellular decisions (Dolmetsch et al., 1997; Kholodenko, 2006; Selimkhanov et al., 2014). Signals, such as extracellular ligand concentrations, are transduced by modulating the enzymatic activities and local concentrations of signaling mediators, such as kinases, within a cell. Traditionally, the unknown structures of signaling networks are reconstructed from many biochemical experiments. They are then formalized as graphs where nodes represent active signaling molecules and directed edges represent interactions between molecules. More recently, statistical modeling has been used to infer network structures in a data-driven manner (Sachs et al., 2005).

In contrast to canonical interpretations of signaling networks as static structures, data-driven approaches of network inference have suggested that the structure of signaling networks may strongly depend on context, including cell phenotype, type of input signal, and treatment, for example, with an inhibitor, as well as a certain degree of additional cell-to-cell variability (Hill et al., 2017; Petsalaki et al., 2015; Will and Helms, 2015; Brightman and Fell, 2000). The context-dependence of signaling is of particular consequence in diseases like cancer, where genetic errors lead to changes in the relative expression or function (Creixell et al., 2015) of signaling proteins and in the local microenvironment, including inputs, that result in signal responses that substantially differ from those of healthy cells (Altschuler and Wu, 2010). Consideration of signaling network rewiring between contexts has led to novel treatment regimens in tumor model systems (Lee et al., 2012). While useful, data-driven network inference requires large quantities of data and some prior knowledge to elucidate causative (i.e., directed) relationships, and these demands obviously increase as more contexts are considered. However, even if contexts are considered, the results of any typical network inference remain static representations of dynamic processes and are fundamentally limited in their ability to allow model-driven predictions of cellular dynamics and responses (Kolitz and Lauffenburger, 2012).

Unlike graph-based network models, mechanistic models of signaling capture both the reaction network structure and the temporal dynamics of signaling, with residual noise attributed to natural cell-to-cell variability or stochasticity. These models consist of sets of reaction rate equations in the form of differential equations that describe how each signaling

component changes over time as a function of the others, and all model components and parameters have unique physical interpretations such as concentrations, binding affinities or reaction rates (Aldridge et al., 2006). Properly calibrated mechanistic models can be used to predict the cellular dynamics from a snapshot of cell state and to analyze the consequences of observed or hypothetical alterations in the relative concentrations or activities of components. The drawback of mechanistic models is that their construction relies on detailed prior knowledge of the reaction network structure and on multiple, targeted experiments that permit calibrating the models parameter values (Aldridge et al., 2006; Kholodenko, 2006). The need for detailed prior information presents a great challenge when multiple contexts are to be considered, as in disease, due to clearly observed variations in network function, which suggests the need for context-dependent model calibrations and possibly for different, context-dependent network structures (Halasz et al., 2016; Kholodenko, 2006). At the same time, biochemical reasoning suggests that the reaction structures themselves should actually be fixed: two reactants may or may not present, but the kinetics of their interaction itself should not change, unless other factors or modulators are altered. This reasoning, in turn, suggests an important consequence. Namely, given explicit observations of signaling components and inputs across contexts, an appropriate mechanistic model should be "context-explicit" by reconciling the entire range of context-dependent signaling dynamics without network rewiring. In other words, it should in principle be possible to construct a single mechanistic model that, combined with snapshots of individual cell states, would be able to explain and predict signaling across many phenotypic contexts as they are present, for example, in individual patient samples. In practice, however, the ability of a single mechanistic model to represent signaling across cell phenotypes is often restricted, due to ill-characterized differences in cellular milieu and gaps in knowledge, which are often considered as sources of natural stochasticity.

Here, for the first time, we use multiplexed single-cell data to calibrate a minimalistic mechanistic model that is capable of consolidating differences in signaling dynamics across two distinct cell phenotypes that at first glance appear to mandate context-based network rewiring, namely, cells before and after an epithelial-mesenchymal transition (EMT). This

developmental program enables polarized epithelial cells to de-differentiate into a dramatically different mesenchymal phenotype that is characterized by loss of cell-cell adhesion junctions, increased capacities for migration and invasion, and resistance to apoptosis (Fu et al., 2018). EMT has been implicated in the generation of metastatic and resistant cancer cell populations, and studies using both bulk (Desai et al., 2015) and single-cell data (Krishnaswamy et al., 2018) have demonstrated EMT-associated alterations in signaling. How these alterations are implemented by the cell is so far unclear, but we demonstrate here that they can be captured by a single mechanistic model. Such a model spanning the entirety of EMT can become a potent tool for understanding and possibly manipulating signaling responses across this critical transition.

To generate both epithelial and mesenchymal cell populations, we use the prviously established and robust experimental model of TGF-β-inducible EMT in Py2T murine breast cancer cells (Waldmeier et al., 2012). We then use mass cytometry to quantify simultaneously the epithelial or mesenchymal cell phenotype as well as both the total expression and phosphorylation dynamics of multiple MAPK/ERK and PI3K/AKT signaling pathway components in response to a typical proliferative signal, the stimulation with epithelial growth factor (EGF). This combination of data permits us not only to develop a context-explicit mechanistic model but also to compare our results directly with the current state of the art. Thus, we begin by applying a classical network-inference method to the data, which seems to suggest quite clearly that the network structure of the ERK and AKT pathways is rewired in a phenotype-dependent manner. Then, using the same data and a mechanistic single-cell modeling approach (Chapter 3) that is solidly based on the core principles of biochemical systems modeling (Savageau, 1976; Voit, 2000, 2013), we construct, *ab initio*, two models of the signaling pathway, one each for the mesenchymal and epithelial cell phenotypes. Intriguingly, the results show that accounting for unmeasured contextual variables allows us to consolidate the two mechanistic epithelial and mesenchymal models into a single model with constant reaction structure, minimal changes in biochemical parameters, and very modest residual noise. This result presents proof-of-principle that no true rewiring is necessary across EMT but that alterations in signaling processes during

this dramatic transition in cell phenotype are instead the result of changes in the relative concentrations of signaling components that can be captured with a single mechanistic model. More generally, our results suggest that further extending this type of single-cell model development, combined with highly-multiplexed single-cell measurements of intra- and extracellular states, has the potential of greatly reducing uncertainty that is otherwise attributed to natural stochasticity, and improving our ability to predict and analyze signaling responses in heterogeneous tissues and different disease contexts.

4.3 Results

The starting question for our analysis was the following: given that cells undergo a dramatic change in phenotype that clearly includes signaling, must the signaling network be rewired or are changes in the relative concentrations of signaling components alone sufficient to explain the altered dynamics (Figure 4.1)? Specifically, we posed the hypothesis that a single, mechanistic model with fixed reaction network structure and kinetic parameters, calibrated with highly-multiplexed measurements of signaling protein state and expression, should be able to reconcile the differences in signaling across phenotypes.

To test our hypothesis, we used an experimental model of EMT and focused on EGF signaling dynamics in the MAPK/ERK and PI3K/AKT pathways, which are considered proliferative, pro-growth and pro-oncogenic in response to epidermal growth factor receptor (EGFR) input (Wee and Wang, 2017). We used mass cytometry to characterize individual cell phenotype (epithelial or mesenchymal) as well as the levels of phosphorylated and total signaling proteins. Specifically, we measured total and phosphorylated forms of Mek, Erk and p90Rsk in the core Erk pathway, Akt and Gsk3β in the Akt pathway and ribosomal protein S6 (S6), a downstream target of both Erk and Akt signaling, in 12 serial samples from a time course of EGF stimulation (for full antibody panel, see Table C.4). In total, our data set contains 31-dimensional measurements across 34 conditions: 13 time points for EGF stimulation and 4 time points for unstimulated control EGF, both made at 2 EMT time points; all generated in triplicate with an average of nearly 10,000 cells per sample.

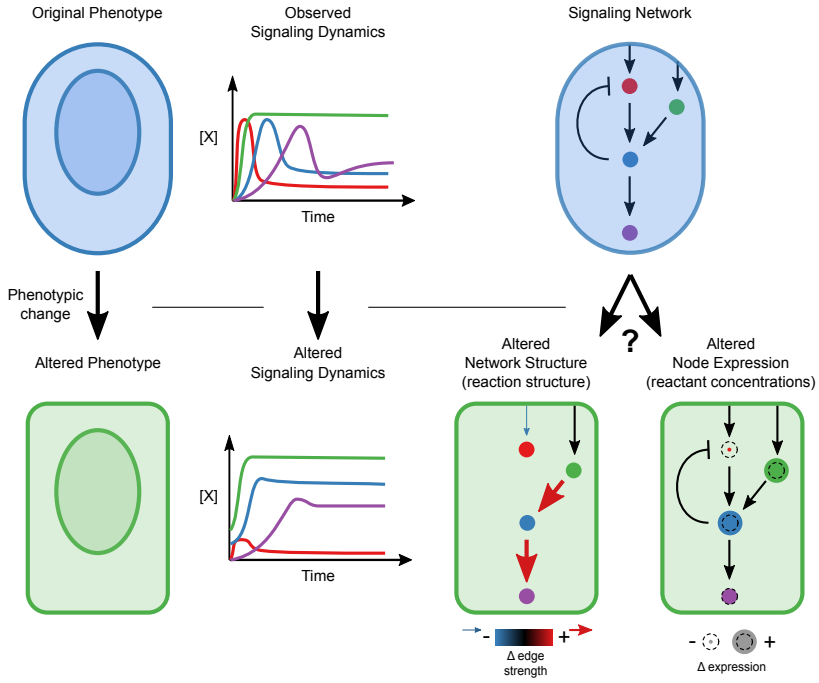
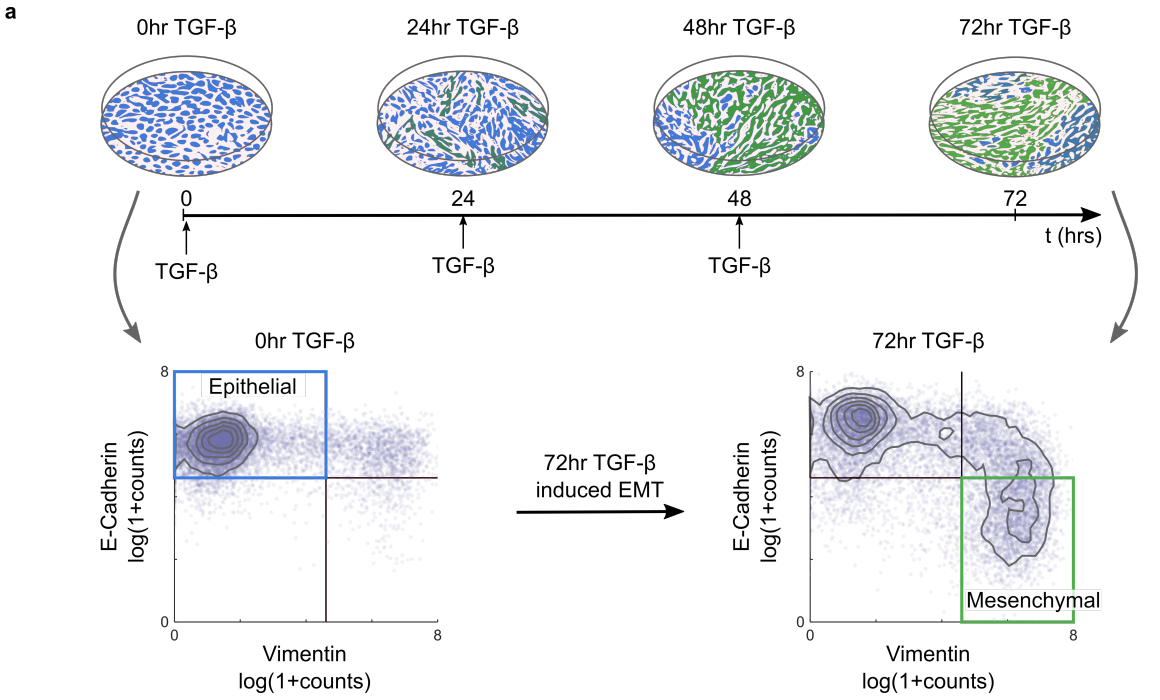
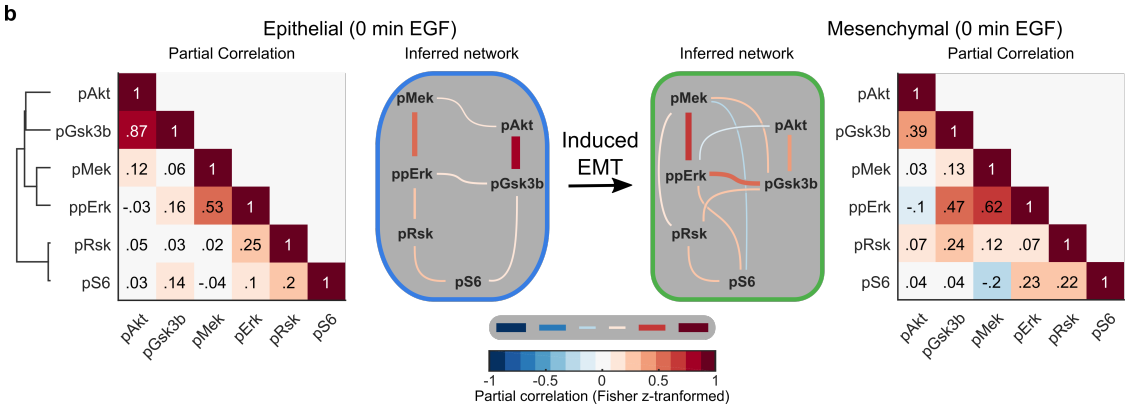


Figure 4.1: Conceptual overview. A phenotypic transition is associated with a changed signaling response. A key question is whether this alteration requires network rewiring, with the addition or elimination of edges, or whether the network structure can be constant, while moderate differences in relative concentrations are sufficient to explain the differences.

4.3.1 Data-driven statistical network inference suggests that the AKT/ERK signaling network is rewired in response to TGF-β treatment

To study signaling during an EMT, we used a model of TGF-β induced EMT in Py2T murine breast cancer cells (Waldmeier et al., 2012; Krishnaswamy et al., 2018). Cells are normally epithelial in visual and molecular phenotype; growing in monolayer with a homogeneous cobblestone appearance and high expression of the epithelial marker E-cadherin. Chronic treatment of cells with TGF-β causes cells to transition to a mesenchymal phenotype that no longer grows in monolayer; it is defined by an elongated shape, loss of E-cadherin, and acquisition of mesenchymal markers such as vimentin. As in previous studies, we defined epithelial cells as E-cadherin-high, vimentin-low in samples untreated with TGF-β (day 0 samples). To generate a population of mesenchymal cells, samples were treated with TGF-β at 24 hour intervals for three days (day 3 samples) and then selected as E-cadherin low, vimentin high (Figure 4.2a; Figure C.1).


a
0hr TGF-β
24hr TGF-β
48hr TGF-β
72hr TGF-β
0
24
48
72
t (hrs)
TGF-β
TGF-β
TGF-β
0hr TGF-β
Epithelial
E-Cadherin log(1+counts)
Vimentin log(1+counts)
72hr TGF-β induced EMT
72hr TGF-β
Mesenchymal
E-Cadherin log(1+counts)
Vimentin log(1+counts)


b
Epithelial (0 min EGF)
Partial Correlation
Inferred network
Induced EMT
Inferred network
Mesenchymal (0 min EGF)
Partial Correlation
pAkt
pGsk3b
pMek
ppErk
pRsk
pS6
Partial correlation (Fisher z-tranformed)

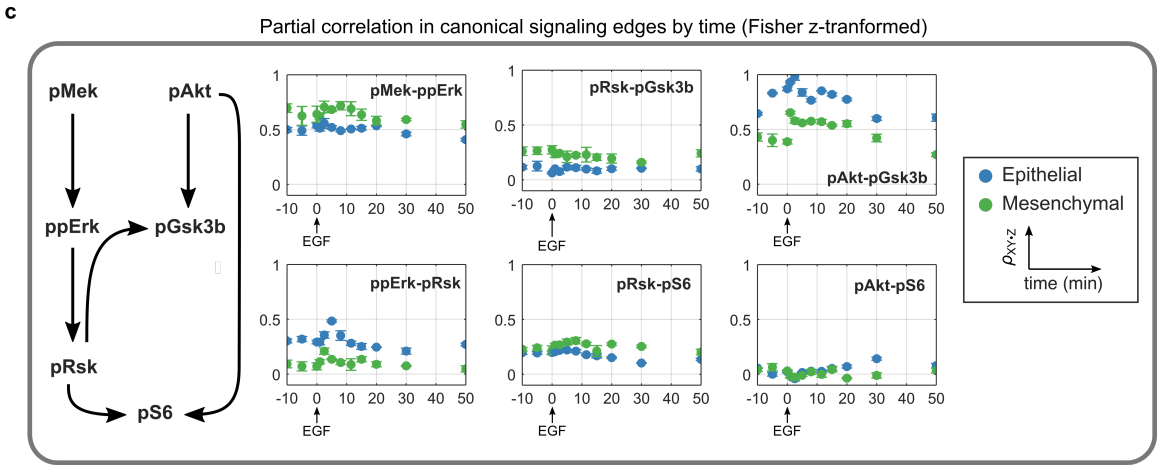

c
Partial correlation in canonical signaling edges by time (Fisher z-tranformed)
pMek
pAkt
ppErk
pGsk3b
pRsk
pS6
pMek-ppErk
pRsk-pGsk3b
pAkt-pGsk3b
ppErk-pRsk
pRsk-pS6
pAkt-pS6
EGF
Epithelial
Mesenchymal
time (min)

Figure 4.2: Partial correlation-based network inference suggests that the ERK/AKT signaling network is significantly rewired during EMT. (a) Overview of an EMT model (top) including an illustration of the visual phenotype and a molecular definition of epithelial and mesenchymal cells measured by mass cytometry at various time points (bottom). **(b)** Partial correlation ($\rho_{XY \cdot Z}$) analysis of phosphoproteins is used to infer the different network structures in epithelial and mesenchymal cells sampled across replicates. The threshold for accepting an edge between X and Y is defined as $\rho_{XY \cdot Z} \geq |0.1|$. All values are Fisher z-transformed. **(c)** Time dependence of signaling strength in canonical signaling edges. Blue and green represent epithelial and mesenchymal phenotypes, respectively. Error bars represent standard deviation of partial correlation across replicates.

To assess whether statistical network inference would suggest that the ERK/AKT signaling network becomes rewired during EMT, we employed the commonly used data-driven approach of partial correlation analysis (Garmaroudi et al., 2010; Desai et al., 2015). The partial correlation between all pairs of phosphoproteins in our panel for epithelial and mesenchymal cells is shown in Figure 4.2b (see Figure C.2 for the analysis including both phospho- and total proteins). Partial correlation values with magnitudes greater than 0.1 (see Methods) were taken to define edges representing true interactions. In epithelial cells, partial correlation recovered the canonical pMek-ppErk-pRsk-pS6 and pAkt-pGsk3β pathways (Mendoza et al., 2011; Manning and Cantley, 2007; Olayioye et al., 2000; Wee and Wang, 2017); note that GSK3β phosphorylation at Serine9 is inhibitory. By contrast, the canonical pAkt-pS6 (via p70S6 kinase) (Mendoza et al., 2011) and pRsk-pGsk3β (Manning and Cantley, 2007) relationships were apparently attributed to pGsk3β-pS6 and ppErk-pGsk3β, respectively. Crosstalk between pMek and pAkt has not been reported, but could indirectly reflect known upstream crosstalk between PI3K3CA and the Raf activator Rac (Ebi et al., 2013).

According to this analysis, EMT clearly appeared to have rewired the signaling network. With the chosen cut-off, the mesenchymal cells appear to lose three (of seven) original edges and gain six new edges. Compared to epithelial cells, the edges ppErk-pRsk, pGskβ-pS6 and pMek-pAkt are lost, while new edges pMek-pRsk, pMek-pGskβ, pMek-pS6, ppErk-pAkt, ppErk-pS6 and pRsk-pGskβ are gained. Noticeably absent in mesenchymal cells is a direct path from the AKT pathway to pS6. For different cut-offs, the results would change somewhat (see Figure C.3), but significant EMT-dependent rewiring would be declared

necessary in all cases.

Each apparent network wiring is a static representation of a signaling process that is in truth dynamic. To assess this time dependence, we measured the signaling dynamics in response to EGF stimulation. Specifically, we quantified the dynamics of network relationships by calculating the partial correlation between widely documented canonical network edges for each time point (Figure 4.2c). This analysis demonstrated a qualitatively different dynamics of the pMek-ppErk edge, as well as differential strengths in most edge relationships, when comparing cells before and after EMT. Notably, EGF stimulation revealed at certain time points two canonical edges that were missed at the steady state: the ppErk-pRsk edge in mesenchymal cells and the pAkt-pS6 edge in epithelial cells. Taken together, the dependence of the network structure on context, such as stimulation, time and cell phenotype, illustrates that a purely statistical analysis, even if it is based on distributions of single cell data, can be misleading as a tool for predicting signaling responses. We will see next that mechanistic modeling with finer resolution remedies this shortcoming.

4.3.2 Mechanistic model with constant network structure represents heterogeneity of epithelial and mesenchymal ERK/AKT signaling

We hypothesized that a mechanistic model with constant reaction network structure, but context- (e.g., phenotype-) dependent parameters, should be able to fit the measured signaling dynamics in both epithelial and mesenchymal phenotypes. Thus, using prior knowledge from the literature, we constructed a reaction network model of the Raf-Mek-Erk-Rsk-S6 and PI3K-Akt-Gsk3β,S6 pathways (Mendoza et al., 2011; Manning and Cantley, 2007; Olayioye et al., 2000; Wee and Wang, 2017) that included pathway crosstalk at the level of PI3K to Rac (Ebi et al., 2013) and RSK to Gsk3β (Manning and Cantley, 2007) (summarized in Figure 4.3a). The reactions were modeled in canonical format (Savageau, 1976; Voit, 2000, 2013) to minimize the inclusion of unmeasured reaction components, which is a streamlining step that has worked well in many other contexts. Most notably, upstream signaling components were not observed, causing us to aggregate all upstream components into a single input with time delay (τ) and an unmeasured modifier variable, pRaf or PI3K, for each pathway. As upstream components were not explicitly modeled (since they could not

be obtained from, or even constrained by, data), it was immediately clear that changes in these components, such as receptor expression and activation state, or their regulation (e.g. altered membrane dynamics) across EMT, had to be represented by changes in parameters, i.e., as magnitudes of pathway inputs.

In contrast to classical model calibration, which targets population averages for parameter estimation, we used a recent approach that uses single-cell data to explicitly model single-cell variation and better constrain model parameters (Chapter 3). This approach simulates sets of individual cells, where the initial state of each cell is taken from a snapshot measurement, and which, when taken together, represent an empirical multivariate distribution analogous to those derived from multiplexed single-cell measurements (see Methods). After model fitting, simulations of independently subsampled cells showed strong agreement between model and data in both the marginal densities and parametric statistical features such as the mean and covariance (Figure 4.3b-c, Figure C.4). Upon inspection, the shape of distribution fits in mesenchymal cells is overall very good; it departs furthest from measurements in the density of pAkt (Figure 4.3b), which may be explained by some additional reaction components or modulators that change during EMT; an example could consist of additional membrane-level variances due to cellular morphological switches. Overall, this result confirms that a constant reaction structure is sufficient to represent signaling across EMT.

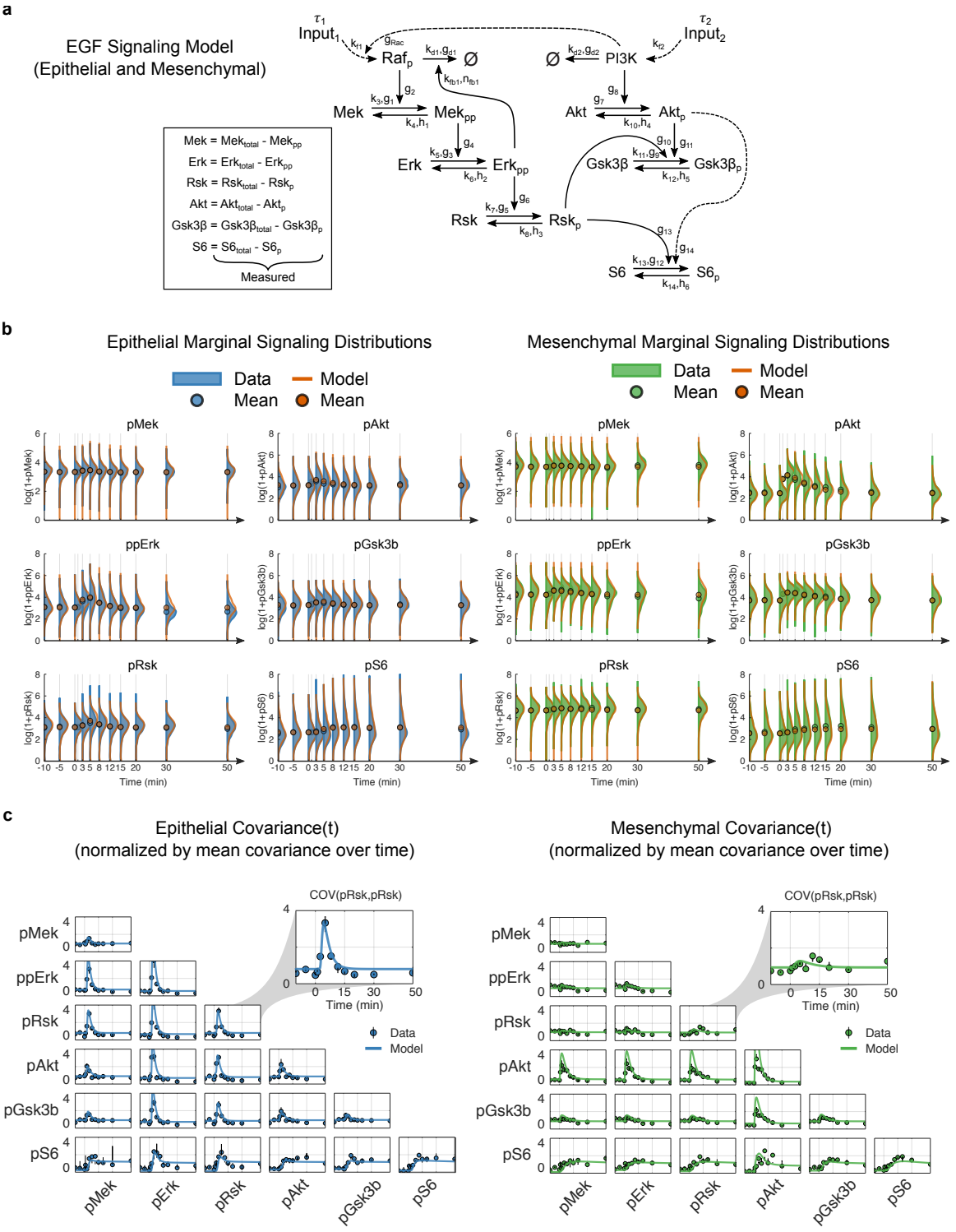

a
EGF Signaling Model
(Epithelial and Mesenchymal)
Mek = Mek_total - Mek_pp
Erk = Erk_total - Erk_pp
Rsk = Rsk_total - Rsk_p
Akt = Akt_total - Akt_p
Gsk3β = Gsk3β_total - Gsk3β_p
S6 = S6_total - S6_p
Measured
b
Epithelial Marginal Signaling Distributions
Mesenchymal Marginal Signaling Distributions
Data
Model
Mean
Mean
pMek
pAkt
ppErk
pGsk3b
pRsk
pS6
Time (min)
c
Epithelial Covariance(t)
(normalized by mean covariance over time)
Mesenchymal Covariance(t)
(normalized by mean covariance over time)
COV(pRsk,pRsk)
Time (min)
Data
Model
pMek
pErk
pRsk
pAkt
pGsk3b
pS6

Figure 4.3: Mechanistic model of EGF signaling in ERK and AKT pathways with parameter fits to epithelial and mesenchymal cells. (a) Model reaction structure and measured variables used for both cell types, annotated with kinetic parameters. **(b)** Marginal distributions of ($\log_2$ transformed) data (solid) and model simulations (orange line) for dynamic signaling variables in epithelial (blue solid, left) and mesenchymal (green solid, right) cells. Circles represent means. Marginal distribution at time t = 1 [min] not shown for improved visual purposes (mean shown). **(c)** Data (symbols) and model simulations (line) of covariance between signaling variables over time in both cell types. Black bars represent the range of covariances across replicates. All values in (b,c) were calculated by subsampling cells across experimental replicates.

4.3.3 Consolidation of ERK/AKT signaling in epithelial and mesenchymal cells requires only minimal adjustments of mechanistic model parameters

Using our unifying reaction network structure, which has the potential of representing the signaling dynamics in both cell phenotypes, we investigated to what extent the reaction parameters could also be held constant across epithelial and mesenchymal cells. The original closest-fit point estimates for the parameter sets differed with each cell type. However, both cases admitted ranges of parameter values within which the data were fit very well, and in most cases, these ranges overlapped between epithelial and mesenchymal cell models. This quality of fit can be gleaned from a grid-based sensitivity analysis of how individual parameter changes alter the quality of a model fit (Figure 4.4a). While these results are encouraging, univariate sensitivity analyses do not account for the potentially complex effects of changing multiple parameters simultaneously. We therefore searched simultaneously for complete, close-by parameter sets that minimized the differences in parameters between epithelial and mesenchymal cells without decreasing model fitness (Figure 4.4b).

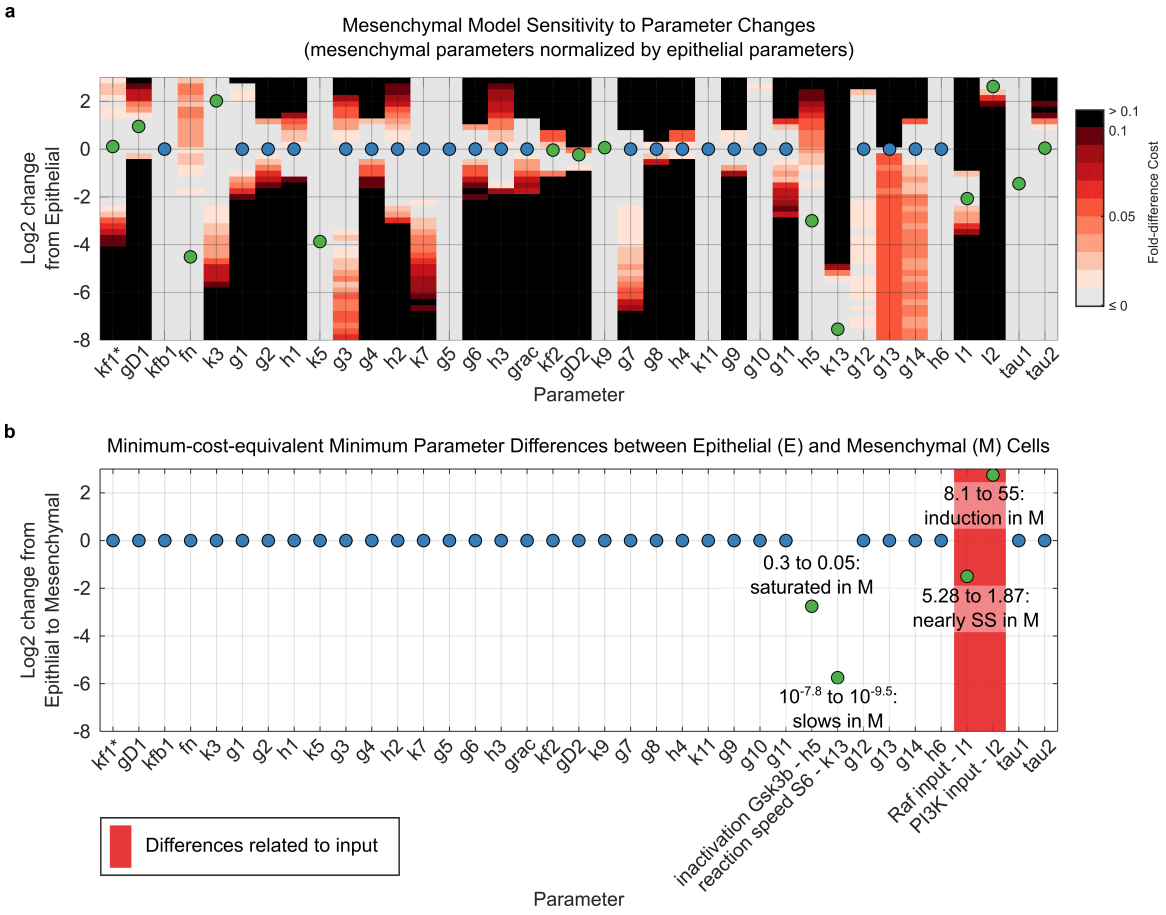


Figure 4.4: Reconciliation of mesenchymal and epithelial model parameters. **(a)** Sensitivity analysis of mesenchymal model parameter perturbations. Grid represents parameter changes on $\log_2$ scale of epithelial parameters used in Figure 3. Circles represent mesenchymal parameters used in Figure 3 (normalized to epithelial parameters in $\log_2$ scale). If epithelial and mesenchymal parameters are equal, the circles are blue; otherwise they are green. Grid color (color bar) represents percent increase in cost function for the mesenchymal cell model given a corresponding parameter change. **(b)** Minimal parameter difference between epithelial and mesenchymal parameter sets for constant mesenchymal function cost. Red highlights those parameters associated with model inputs (input magnitude or degradation). Scale and circle colors as in (a).

As we had hypothesized, based on generic biochemical reasoning, analysis of the mechanistic model revealed that only four out of 38 parameters require adjustments in values during the transition from the epithelial (E) to the mesenchymal (M) state: First, I_1, the magnitude of the input to the ERK pathway, decreases from 5.3 (E) to 1.9 (M), approaching the steady-state value of 1. Second, I_2, the magnitude of the input to the AKT pathway, increases by less than an order of magnitude from 8.1 (E) to 55 (M). Third, parameter h_5 decreases from 0.3 (E) to 0.05 (M). This parameter is related to the sensitivity of pGsk3β

dephosphorylation in response to increases in pGsk3β, and the decrease actually causes just a minute 1% change in model quality (fitness). This result may not be surprising given that (1) this parameter value has the direct interpretation of the reaction moving from near zero-order (nearly saturated with respect to pGsk3β) to very near zero-order (even closer to saturation), and (2) pGsk3β is generally higher in (M) as in (E), both at the steady state and at the maximum. Finally, k_{13}, which is related to the maximal rate of S6 phosphorylation, decreases from $10^{-7.8}$ (E) to $10^{-9.5}$ (M). This change is less clearly explained, but could be related to altered expression of missing reaction components or differences in local concentrations or diffusion rates. Taken together, almost all of the systems responses are explained without evoking natural stochasticity.

As an independent model validation, we compared our findings with results of (Salt et al., 2014) showing downregulation of ERBB3, an EGF receptor related to ERK pathway activation, and upregulation of PI3KCA, which encodes the catalytic p110α subunit of PI3K, during EMT. The results of these studies agree well with model predictions of decreases or increases in ERK and AKT pathway inputs, respectively, which, as discussed before, are all subsumed in the model inputs, due to insufficient information. This consistency with independent observations supports our EMT-spanning model and provides clear direction for additional variables that must be measured for a complete reconciliation of the remaining model parameters across EMT.

4.4 Discussion

From a theoretical biochemical perspective, differences in signaling responses across cells should be determined by relative differences in concentrations and states of reaction components, rather than by substantial reaction network rewiring. We show, for the first time to our knowledge, that this conjecture indeed applies to signaling in cells before and after a phenotypic transition, such as EMT. To demonstrate this consistency, we generated very informative data. Specifically, we measured twice as many state variables as had been used to date to characterize the dynamics of a single-cell- or distribution-level mechanistic model of signaling (Chapter 3). Given the ability of calibrated mechanistic models to analyze,

simulate, and explain the dynamics from an initial snapshot of cell states, it was possible to initialize an appropriate model – or set of models – with single-cell measurements from, for example, a tumor sample that contains multiple cell types. The proof-of-principle we present here represents a large and important progression towards construction of such tumor-level models.

To assess the conjecture of a single network structure encompassing different contexts, we combined an *in vitro* model of EMT with multiplexed single-cell measurements and with computational modeling to determine how EGF signaling in the ERK and AKT pathways changes with cell phenotype. This approach included the generation of a rich single-cell data set that is especially valuable for use with mechanistic modeling approaches due to the uncommon inclusion of total markers that constrain variation in expression. For fair comparisons, we first used a traditional data-driven approach of network inference based on partial correlation, which suggested that the ERK/AKT signaling network had to be rewired in a phenotype-dependent manner during EMT, thereby reflecting conclusions published in the literature. In stark contrast, a more appropriate dynamic, mechanistic model with constant network structure and near constant kinetic parameter values, calibrated with state-of-the-art single-cell data, does not need rewiring: it is able to reconcile the variation in signaling dynamics in both epithelial cells mesenchymal cells across EMT with minor cellular adjustments. In other words, our results provide clear evidence that a properly calibrated mechanistic model can represent signaling across a contextual change as large as EMT, and that actual cells apparently do not need to rewire their signaling networks but simply modulate their reaction components in a minor fashion to alter signaling.

While the ability to explain variation in signaling responses from an individual snapshot of cell states is an exciting prospect, we emphasize the necessary considerations to apply our approach. Primarily, these are the necessity to explicitly measure and model the primary sources of variation in system components that affect the signaling dynamics, which could include not only levels or functional state of signaling proteins, but also other features including the cells microenvironment. Thus, the main limitations are directly related to the ability of measuring the appropriate variables, as well as the computational cost of

fitting mechanistic models to increasingly large systems. These limitations, however, continually diminish as multiplexed single-cell measurement technologies and computational infrastructure rapidly improve.

Taken together, a single mechanistic model with entirely constant reaction structure at the single-cell level can reconcile the EGF signaling dynamics of the ERK/AKT signaling network across an EMT. The only differences, which are without doubt associated with the different signaling responses of the system at E and at M, are modest changes in the magnitudes of pathway inputs and slight alterations in only two of the models 34 kinetic parameters. This result clearly demonstrates that actual murine Py2T breast cancer cells have no need for rewiring the reaction structure of their signaling network during EMT. No mechanisms or processes have to be added and none have to be removed. In fact, it appears that the signaling system is essentially deterministic and that the formerly assumed stochasticity is by and large a matter of unmeasured variables and a modest degree of intercellular variability.

4.5 Methods

Cell culture

Py2T cells were obtained from the laboratory of Gerhard Cristofori, University of Basel, Switzerland (Waldmeier et al., 2012). Cells were tested for mycoplasma contamination upon arrival and regularly during culturing and before being used for experiments. Cells were cultured at 37°C in DMEM (D5671, Sigma Aldrich), supplemented with 10% FBS, 2 mM L-glutamine, 100 U/ml penicillin, and 100 μg/ml streptomycin, at 5% CO_2. For cell passaging, cells were incubated with TrypLE Select 10X (Life Technologies) in PBS in a 1:5 ratio (v/v) for 10 minutes at 37°C. For each experiment, cells were seeded at the density of 0.3 million cells per plate (100 mm diameter) and allowed to recover for 36 hours.

TGF-β stimulation

After reaching 60% confluence, cells were either mock treated or treated with 4ng/ml TGF-β (Human recombinant TGF-β1, Cell Signaling Technologies) for 72 hours. Cell growth medium and 4ng/ml TGF-β treatment were renewed every 24 hours until 24 hours before

any EGF stimulation experiments. For each condition, three biological replicates were cultured, harvested and analyzed.

Cell harvesting and EGF stimulation time course

For cell harvest, cells were washed two times with PBS and incubated with TrypLE Select 10X (Life Technologies) in PBS at a 1:5 ratio (v/v) for 10 minutes at 37°C. Following cell detachment, cells were mixed and resuspended in serum-free media and allowed to recover from detachment for two hours at 37°C and 5% CO_2 with periodic shaking to avoid cluster formation. After the recovery period, samples were taken to establish baselines. EGF (Peprotech) was then added to a final concentration of 100 ng/ml and sampling continued to characterize signaling dynamics. Samples were taken at {-10,-5,0,1,3,5,8,12,15,20,30,50} minute time points relative to stimulation ($t = 0$) with EGF (the 0-minute sample was not stimulated). At the time of sampling, paraformaldehyde (PFA, from Electron Microscopy Sciences) was added to the cell suspension to a final percentage of 1.6%, and cells were incubated at room temperature for 10 minutes. Crosslinked cells were washed twice with cell staining medium (CSM, PBS with 0.5% BSA, 0.02% NaN_3) and after centrifugation, ice-cold methanol was used to resuspend the cells, followed by a 10-minute permeabilization on ice or for long-term storage at -80°C.

Metal-labeled antibody conjugation

The MaxPAR antibody conjugation kit (Fluidigm) was used to generate isotope-labeled antibodies using the manufacturer's standard protocol. After conjugation, the antibody yield was determined based on absorbance at 280 nm. Candor PBS Antibody Stabilization solution (Candor Bioscience GmbH) was used to dilute antibodies for long-term storage at 4°C.

Barcoding and staining protocol

Formalin-crosslinked and methanol-permeabilized cells were washed three times with CSM and once with PBS. Cells were incubated in PBS containing barcoding reagents (^{102}Pd, ^{104}Pd, ^{105}Pd, ^{106}Pd, ^{108}Pd, ^{110}Pd, ^{113}In and ^{115}In) at a final concentration of 100 nM for 30

minutes at room temperature and then washed three times with CSM (Bodenmiller et al., 2012). Barcoded cells were then pooled and stained with the metal-conjugated antibody mix (Table C.4) at room temperature for 1 hour. The antibody mix was removed by washing cells three times with CSM and once with PBS. For DNA staining, iridium-containing intercalator (Fluidigm) diluted in PBS with 1.6% PFA was incubated with the cells at 4°C overnight. On the day of the measurement, the intercalator solution was removed, and cells were washed with CSM, PBS, and ddH_2O. After the last washing step, cells were resuspended in ddH_2O and filtered through a 70-μm strainer.

Mass cytometry analysis

EQ Four Element Calibration Beads (Fluidigm) were added to cell suspensions in a 1:10 ratio (v/v). Samples were analyzed on a Helios mass cytometer (Fluidigm). The manufacturer's standard operation procedures were used for acquisition at a cell rate of approximately 500 cells per second. After the acquisition, all FCS files from the same barcoded sample were concatenated (Bodenmiller et al., 2012). Data were then normalized, and bead events were removed (Finck et al., 2013) before doublet removal and de-barcoding of cells into their corresponding wells using a doublet-filtering scheme and single-cell deconvolution algorithm (Zunder et al., 2015). Cytobank (http://www.cytobank.org/) was used for additional gating on the DNA channels (^{191}Ir and ^{193}Ir) and ^{139}La/^{141}Pr to remove remaining doublets, debris and contaminating particulates. Data were then exported as .fcs files for subsequent analysis.

Gating epithelial and mesenchymal cells

Epithelial were gated as E-cadherin high / vimentin low in samples without TGF-β treatment. Mesenchymal cells were gated as E-cadherin low / vimentin high in samples treated for three days with TGF-β. Gating cut-offs are shown in Figure 4.2a. Treatment with TGF-β for longer than three days increases the percentage of Py2T population that undergoes EMT. Samples treated with TGF-β for three days also contained cells within the "E-cadherin high / vimentin low" gate, but these were not considered contextually as epithelial cells.

Data normalization and scaling for use in modeling

Experimental measurements were normalized for comparisons across independent experiments and measurements. Cell events with greater than 20 counts for cleaved PARP were removed as dead or apoptotic. Before using variables in modeling, they were linearly scaled to satisfy biological constraints, such as the total units of a protein must be greater than or equal to the units of the phosphorylated form (see Appendix C). Cells used in fitting were subsampled across experimental replicates to reduce computational cost. A full description of normalization, scaling and subsampling of data before use in modeling is presented in Appendix C.

Partial correlation-based network inference

Given two random variables X and Y and a set of controlling variables $\mathbf{Z} = Z_1, ..., Z_n$, the partial correlation $\rho_{XY \cdot \mathbf{Z}}$ is a measure of the relationship between X and Y when the effects of the $\mathbf{Z} = Z_1, ..., Z_n$ random variables have been accounted for. Mathematically, $\rho_{XY \cdot \mathbf{Z}}$ is the correlation of the residuals e_X and e_Y that result from a linear regression of X and Y with $\mathbf{Z}$, respectively.

To determine the cutoff for partial correlation-based network representations, thresholds can be used to define a minimum p-value or correlation coefficient. In order to focus on stronger relationships, we used a threshold on the partial correlation values. The choice of $\rho_{XY \cdot \mathbf{Z}} \geq |0.1|$ as the threshold was made as a qualitative boundary between maximizing canonical and minimizing non-canonical signaling relationships in epithelial cells. Most notably, this setting captured the edge between the ERK pathway and pS6, as well as some form of crosstalk between the ERK and AKT pathways. The addition or subtraction of edges based on other thresholds may be readily calculated from the heatmaps provided in Figure 4.2 and, in the context of both total and phosphoproteins, in Figure C.2. Figure C.3 illustrates the relationship between the partial correlation threshold and network edge number. Heatmap labels were ordered by hierarchical clustering the epithelial population values using single-linkage clustering and Euclidean distance.

Mechanistic model of ERK/AKT pathway response to EGF

We used the 'Single-cell ODE modeling' (SCODEM) approach as described in Chapter 3. Briefly, cells are assumed to have the same population-level kinetic parameters, which are determined by minimizing the maximum mean discrepancy (MMD)(Gretton et al., 2012a), a statistical two-sample test of similarity for n-dimensional distributions, between simulated and experimentally measured distributions. Cell-to-cell variation in unmeasured components is captured in a subset of rate constants that are algebraically determined by a combination of model structure, population-level parameters and steady-state measurements. To fit individual parameter sets for epithelial and mesenchymal cell populations, we subsampled 500-1000 cells across the three replicates for each time point. Model equations may be found in Appendix C.

Parameter optimization

Parameter optimization was performed using a combination of global and local search methods. First, 50,000 initial parameter sets were sampled from a user input parameter range. Next, the 200 parameter sets with minimum model cost were selected. Finally, each parameter set was refined using multiple rounds of optimization using the unconstrained local search algorithm `fminsearch` in the Matlab Optimization Toolbox (The MathWorks, Inc.). The `fminsearch` algorithm begins with a broad search perspective before focusing on a more precise local area; a regular re-initialization of the search improved results. Thus, each round of optimization was run for 300 iterations. If at the end of a round, the cost value had improved by at least two percent, another round was initialized using the parameter set output by the algorithm. Otherwise, the optimization was terminated. After this local search approach, `fminsearch` was initialized a final time with the parameter set corresponding to the lowest cost among the 200 solutions and run until the model cost stabilized.

Sensitivity analysis of mechanistic model parameters

A grid-based sensitivity analysis was performed in the context of the population parameters Θ. The cells sampled for model fitting, either epithelial or mesenchymal, and the model structure were held constant. Given the best-fit point estimate of the parameter set for epithelial cells $\Theta_e^* = \theta_{e,1}, ..., \theta_{e,m}$, sensitivities were calculated as the change in cost function F given a change in the j^{th} parameter $\theta_{e,j}$:

$$\frac{\Delta \mathrm{F}}{\Delta \theta_{e,j}}$$

where $\Delta\theta_{e,j}$ was defined using a $\log_2$ fold-change range of [-8:0.25:3] when applied to the best parameter $\theta_{e,j}^*$ for epithelial cells. Notably, this definition implies that parameter sensitivities for the epithelial and mesenchymal models were calculated using the same set of values.

Reconciliation of mechanistic model parameters across EMT

Reconciliation between the best points estimates for epithelial and mesenchymal population parameter sets Θ_e^* and Θ_m^*, respectively, was performed by finding the minimum difference between all corresponding population parameters θ_e and θ_m without increasing the cost F associated with either best-fit parameter set:

$$\min |\Theta_e - \Theta_m|$$

subject to

$$F(\Theta_e) \leq F(\Theta_e^*)$$

$$F(\Theta_m) \leq F(\Theta_m^*)$$

CHAPTER V

CONCLUSION

Cell-to-cell variation has complex and important consequences for intracellular signaling. In many contexts, cells are able to perform essentially identical functions despite their differences, in other contexts, for example in cancer, cell-cell differences in state propagate to differences in function. The overall goal of this dissertation was to develop mathematical and computational methods for the study of cell-to-cell variation in signaling, and to use these tools to increase our understanding of when single cell differences do, or do not, make a meaningful difference. Here, we summarize our scientific contributions towards that goal and discuss future directions related to our work.

5.1 Summary of Results

Chapter 2

In Chapter 2, we established a novel experimental framework that combines mass cytometry with transient transfection to characterize protein expression-dependent effects on signaling systems in a high-throughput manner. We used this framework to measure the effects of independently overexpressing 20 protein kinases on signaling dynamics in the EGFR signaling network for a total of 360 conditions; we measured an average of 11,000 cells per condition with a panel of 35 antibodies that characterized the EGFR network signaling state. To quantify the potentially complex relationships between protein abundance and the state of kinases within the EGFR network in manner suited to these data, we developed a statistical measure called binned-pseudo R-squared (BP-R^2), which identifies many non-monotonic relationships that are not properly captured with correlation metrics such as Spearman correlation. After benchmarking and validation, we used BP-R^2 to identify potentially novel relationships between signaling kinases and cases where protein overexpression altered signaling dynamics in an abundance-dependent manner.

Chapter 3

In Chapter 3, we presented SCODEM: a novel distribution-free computational approach to infer multiplexed signaling trajectories from snapshot single-cell data. Our approach successfully overcame previous experimental limitations that forced a trade-off between continuous and multiplexed measurements. We applied SCODEM to study the consequences of cell-cell variation for how the MAPK/ERK pathway responds to EGF; this included application of the experimental framework developed in chapter 2.

We showed that: (1) cell-to-cell variation in signaling can be described deterministically, given the initial state of each cell; (2) cell volume is a significant contributor to the variability observed in single-cell measurements; (3) signal duration is reliably transmitted through the RAF/MEK/ERK cascade, whereas signal amplitude is a fold-change response to the initial protein state that is uncorrelated across the signaling cascade. More generically, we demonstrated that SCODEM can be a valuable approach to studying reaction kinetics and variability in cellular responses to drug treatments and in disease.

Chapter 4

In Chapter 2, we combined an *in vitro* model of EMT with multiplexed single-cell measurements and mathematical modeling to determine how EGF signaling in the ERK and AKT pathways changes with cell phenotype.

For the important case of an epithelial-mesenchymal transition we showed that signaling network rewiring is not necessary. This result is in stark contrast to the outcomes of traditional data-driven approaches of network inference, which suggest that the ERK/AKT signaling network must be rewired in a phenotype-dependent manner during EMT. To obtain this shift in paradigm, we used SCODEM and showed that an appropriate dynamic, mechanistic model with constant network structure and near constant kinetic parameter values is able to reconcile the variation in signaling dynamics in both epithelial and mesenchymal cells across EMT. This result provides clear evidence that a properly calibrated mechanistic model can represent signaling across a contextual change as large as EMT, and that actual cells apparently do not need to rewire their signaling networks but simply

modulate the levels of their reaction components to alter signaling responses. This result is not due to improved data, since we used exactly the same data with a traditional network inference method, which suggested mandatory rewiring. Instead, it is the combination of powerful single-cell data with an appropriate dynamic model capable of demonstrating that no rewiring is needed.

5.2 Future Directions

Direct experimental validation of single-cell signaling trajectories

We developed the SCODEM framework (Chapter 3) in part because there are no current experimental methods able to measure the state of many signaling components simultaneously and as continuous time series. Thus, direct experimental validation of our model predictions of signaling trajectories in individual cells is currently not possible. Instead, we validated our approach by (1) using methods of algorithm and model construction that are firmly based on accepted assumptions derived from first principles and (2) applying the methods to increasingly demanding examples of cellular heterogeneity, which failed to falsify our results. Nonetheless, a true and more direct validation would be a direct experimental characterization of the accuracy of trajectories in individual cells, which is currently not possible, as there are no live-cell methods to observe what we have modeled, e.g., kinase phosphorylation.

To circumvent this technological issue, we have begun to work in this direction by combining ERK KTR reporters with cutting-edge multiplexed snapshot imaging methods developed in the Pelkmanns group (Gut et al., 2018). Early results indicate that such a combination of methods is possible in principle, and our preliminary data are quite interesting. However, as is to be expected, these experiments have their unique challenges, which will require dedicated methodological effort. While many of these challenges are technical in nature and will be overcome with time, other challenges are fundamental to the use of KTRs specifically, or of genetically encoded sensors in general, and extreme caution will be needed before one can validly accept live-cell readouts as ground-truth. For example,

our preliminary studies indicate an expression-dependent background signal and clear saturation of the KTR cytoplasmic:nuclear ratio (C/N) readout in cells stimulated with EGF at concentrations classically used to characterize signaling dynamics, which immensely restricts the range of useable data. Additionally, preliminary modeling results have shown the potential for the C/N ratio to oscillate even in cases where Erk phosphorylation is perfect or near-perfectly adapted, although it is widely accepted that phospho-Erk levels do not oscillate (Ferrell, 2016). The cause of these artefactual oscillations is not entirely understood but may primarily be due to a combination of limited dynamic range and delays in the reporter system related to nuclear import and export. These issues are not necessarily insurmountable, but illustrate how such reporters introduce new uncertainties into the system that will need to be constrained with innovative, precise experimental designs such as those used to partially characterize the original JNK KTR (Regot et al., 2014).

Further optimization of SCODEM

After developing the SCODEM framework in Chapter 3, our primary focus turned to its application to experimental data as a test of the algorithms underlying assumptions. To this end, we evaluated and optimized the computational performance of SCODEM to the point that suited our needs, which included the ability to run on a single workstation for the problems we considered. However, several directions for improvement remain open. First, and most obvious, is an expansion toward cluster computing. The multistart optimization approach we used is parallelized; it is one of a few processes that could be distributed on a cluster for algorithm speed-up. Second, optimized kernel choice could potentially improve the ability of MMD to discriminate between distributions. Some work has been done on this problem (Gretton et al., 2012b), and it would be interesting to see it tailored to the specific distributions expected in single-cell data. Third, theoretical work could possibly determine optimal sampling of the cell population for model fitting. The most expensive step in the optimization routine is numerically solving the ODE system, which must be performed for each simulated cell. Minimizing the number of cells needed for parameter estimation can clearly speed-up the process. This number will certainly be dependent on the dimension of

the system and the shape characteristics of the state-space distribution. Finally, it would be interesting to investigate some hybrid combination of our framework with sigma-point or generic particle filters as a method of increasing efficiency.

Continuous functional characterization of latent variability

The definition of a cellular phenotype generally involves a discretization of a continuous space of cell states (see, for example Trapnell et al. (2014)). In Chapter 3 we used transient transfection of ERK (as established in Chapter 2) to generate a continuous range of ERK overexpressing cells. Then, rather than discretizing the expression space or relying on mixture models, we found continuous functional representations to describe the effects of ERK overexpression on signaling dynamics. This task was simplified by the controlled nature of the experiment, the direct measurement of ERK abundance and some knowledge of the reaction structures involved. Chapter 4 presented us with a similar, but less well-defined situation. Namely, we know that cells make a continuous transition from an epithelial to a mesenchymal phenotype. As many cell features change during this transition, we discretized this continuous process of EMT into only the well-defined initial and end states and used a model to infer EMT-related changes in variables that were not measured and thus modeled as parameters. An interesting next question therefore is: Can we find a continuous function that maps the change in unmeasured EMT-related variables to progression along an EMT "trajectory"? Several algorithms are now available that attempt to infer developmental trajectories or manifolds in the cellular state space from mass cytometry or single-cell RNA data, for example (Trapnell et al., 2014; Krishnaswamy et al., 2018), and could be used to further this exciting idea.

Application of reaction model-based single-cell trajectory inference in non-signaling systems

Finally, a future project could be the application of the general principles of the SCODEM modeling approach to systems beyond signaling, where single-cell simulations and distribution-free two-sample testing are beneficial. For example, given current single-cell experimental technologies, a promising extension could be in the domain of gene regulatory

networks, perhaps by combining single-cell multiplexed RNA imaging methods, such as MERFISH (Chen et al., 2015), with subsequent protein-level measurements via traditional imaging methods, imaging mass cytometry (Giesen et al., 2014) or genetically encoded fluorescent reporter systems. We would expect simulations of cell ensembles to generate populations for model fitting with multivariate two-sample tests such as MMD to remain an effective approach. However, reaction components in transcription and translation reaction processes are generally present in low copy number and are likely to be more appropriately modeled as stochastic processes, which would require adaptations of our algorithms toward, for example, stochastic differential equations or even agent-based models, as well as necessarily enhanced computational resources. More generally, quantifying the relative contribution of stochastic reaction noise to cellular fates remains an interesting open question that such our methods could, perhaps, be used to address.

5.3 Closing Comments

This dissertation has provided an original framework to infer and analyze biochemical systems at the level of single cells, and resulted in a strong argument that intracellular signaling is primarily a deterministic process at the single cell level. The tools, ideas and results presented in this dissertation are hoped to inspire new avenues of thinking toward understanding cellular decision making across many contexts including disease.

APPENDIX A

SUPPLEMENTARY MATERIALS: INFLUENCE OF NODE ABUNDANCE ON SIGNALING NETWORK STATE AND DYNAMICS ANALYZED BY MASS CYTOMETRY[1]

A.1 Supplementary Figures

[1]Adapted from (Lun et al., 2017).

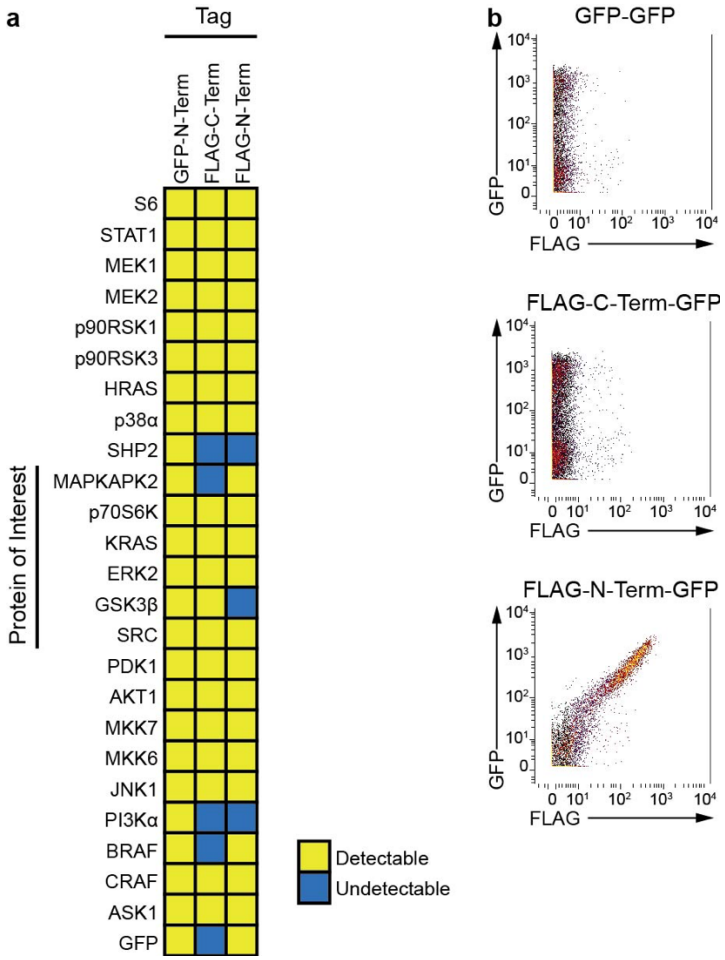


Figure A.1: Technique validation. **(a)** Detection of GFP-N-terminal, FLAG-C-terminal, and FLAG-N-terminal tagged proteins. All GFP-tagged fusion proteins, but only 20 of the 25 FLAG-C-terminal tagged and only 22 of the 25 FLAG-N-terminal tagged proteins, were detected using mass cytometry. **(b)** HEK293T cells overexpressing GFP-GFP, FLAG-C-terminal-GFP, and FLAG-N-terminal-GFP fusion proteins were co-stained with anti-GFP and anti-FLAG antibodies. The fusion protein FLAG-C-terminal-GFP was detected by the anti-GFP antibody but not with the anti-FLAG antibody. This indicates that in certain contexts the FLAG tag is not accessible to the anti-FLAG antibody. The FLAG epitope may be masked due to protein folding or by the denaturation process that is part of our experimental protocol.

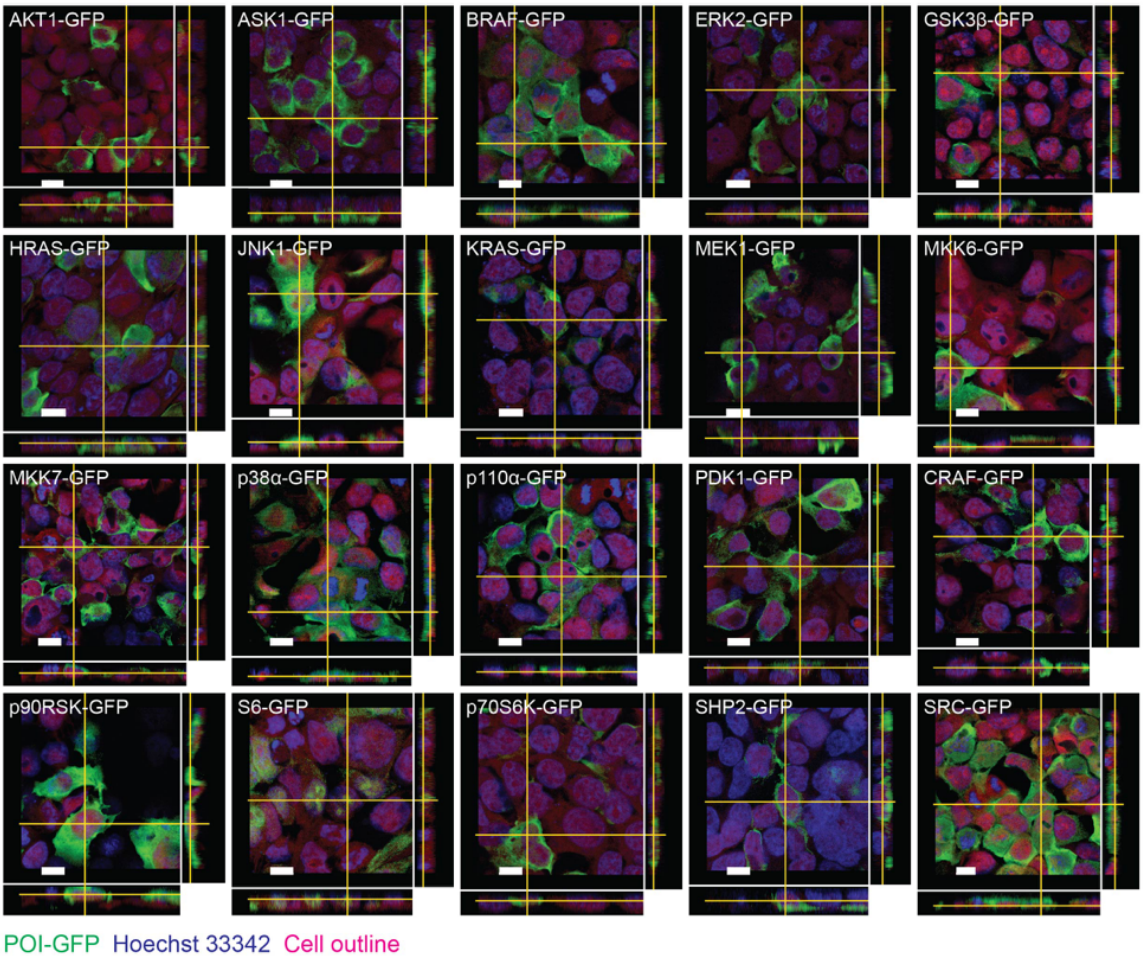


Figure A.2: GFP-tagged POIs have normal localization. HEK293T cells that overexpressed the GFP-tagged POIs used in this study were imaged with confocal microscopy. For each POI, the main panel shows the image in a given z-depth; the bottom panel and the side panel shows x-z and y-z cross-sectional images, respectively. POI-GFP subcellular localization was determined by overlapping with two control stains: Hoechst 33343 for the nucleus and Alexa Fluor 647 carboxylic acid succinimidyl ester indicating the cell outline. The POI-GFP localization was verified by comparison with information of the UniProt subcellular localization database (Supplementary Table 4 online).

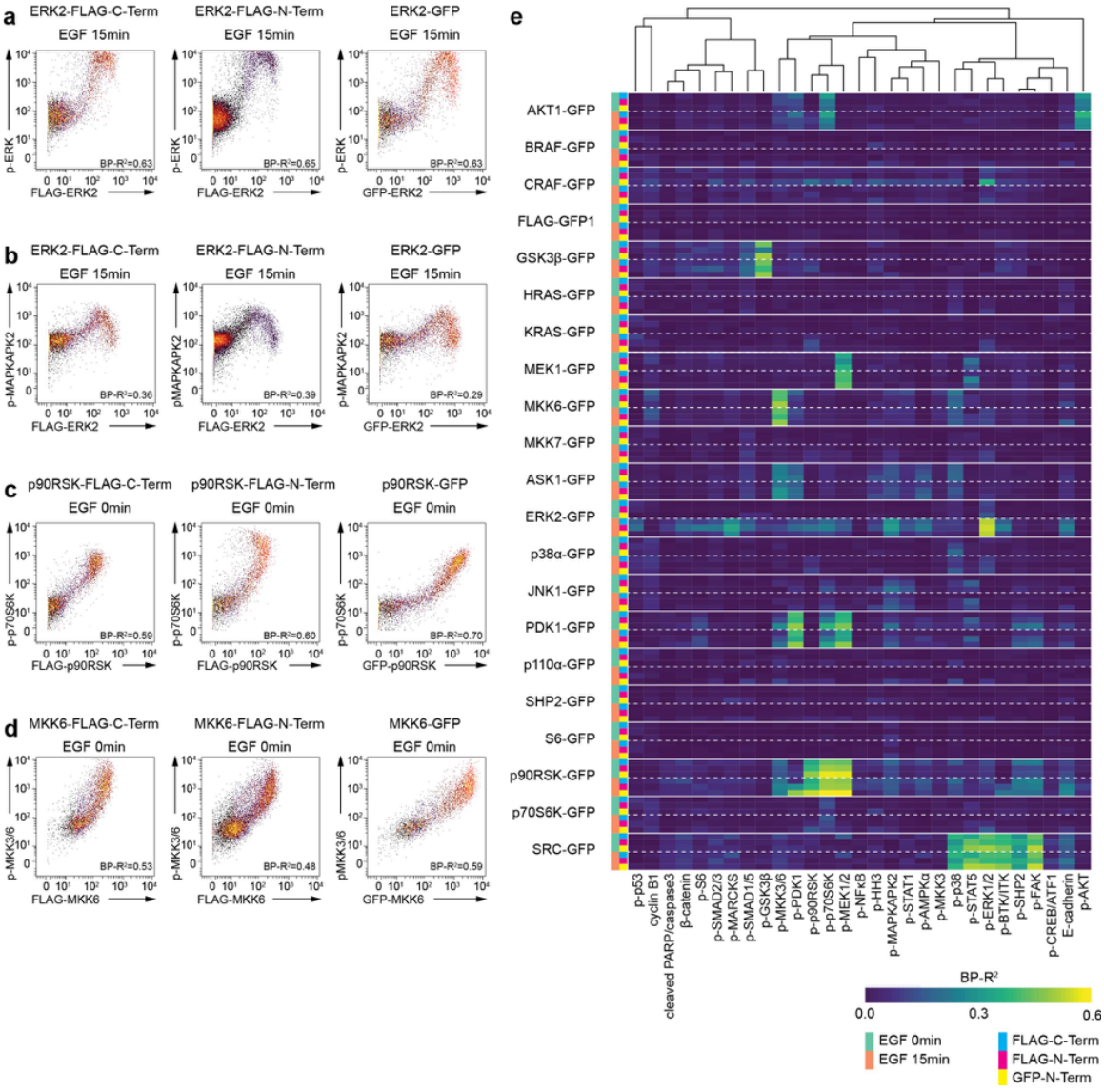


Figure A.3: GFP tag does not disrupt catalytic activities of POIs. (a-d) Catalytic activities of GFP-tagged POIs were compared with FLAG-C-terminal and FLAG-N-terminal tagged POIs. The examples shown here indicate that the GFP tag did not alter signaling relationships or signaling dynamics after EGF stimulation (the complete dataset with comparison of all constructs used in this study is shown in Supplementary File 1 online). **(e)** Heat map showing abundance-dependent signaling relationship strengths from overexpressed POIs with three different tags as determined by BP-R^2 analysis (Figure 2.3 and Methods). Measured markers showing at least one strong relationship in any of the conditions were included in the heat map. Strong relationships were detected independently of tag. BP-R^2 values slightly vary for the 3 tags, due to the antibody accessibility and differences in transfection efficiencies (Figure A.1).

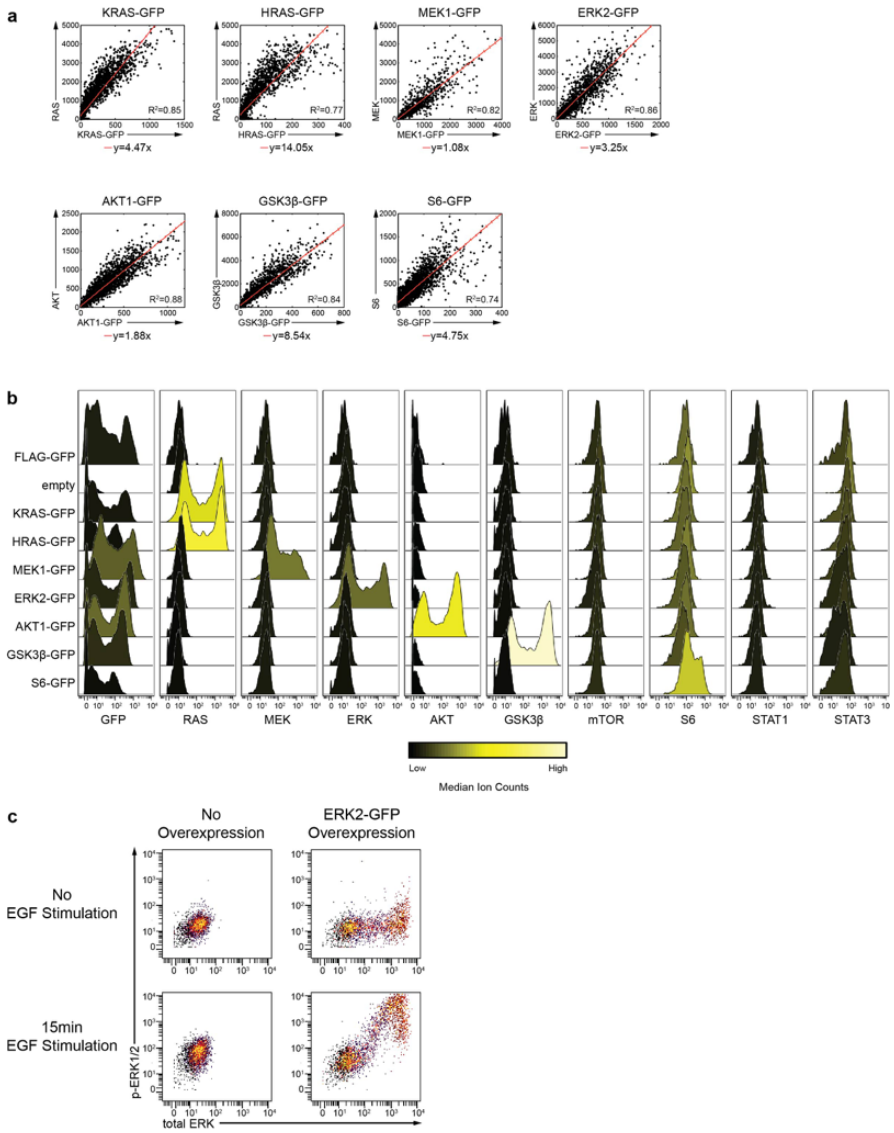


Figure A.4: Total protein antibody staining of HEK293T cells overexpressing a GFP-tagged POI. (a) HEK293T cells transfected with KRAS-GFP, HRAS-GFP, MEK1-GFP, ERK2-GFP, AKT1-GFP, GSK3β-GFP, or S6-GFP for 18 h were stained with anti-total POI and anti-GFP antibodies. A linear regression analysis for each pair was performed in the original scale. R^2 ranges from 0.74 to 0.88, indicating the total POI is linearly correlated with GFP and that the POI overexpression does not alter the expression of the endogenous POI. **(b)** The same cells were stained with nine antibodies to quantify total protein as well as with a GFP antibody. Median ion counts for all measured markers are shown. Overexpression of a POI-GFP for 18 h does not cause notable changes in the measured network nodes. **(c)** ERK2-GFP transfected HEK293T cells and the untransfected control with or without EGF stimulation were stained for total-ERK and phospho-ERK (Thr202/Tyr204). The dynamic range in the overexpression condition allows observation of abundance-dependent signaling relationships. With total ERK staining, the same signaling relationships as shown in the Supplementary Figure A.2b is recapitulated, verifying GFP as an indicator of POI expression level.

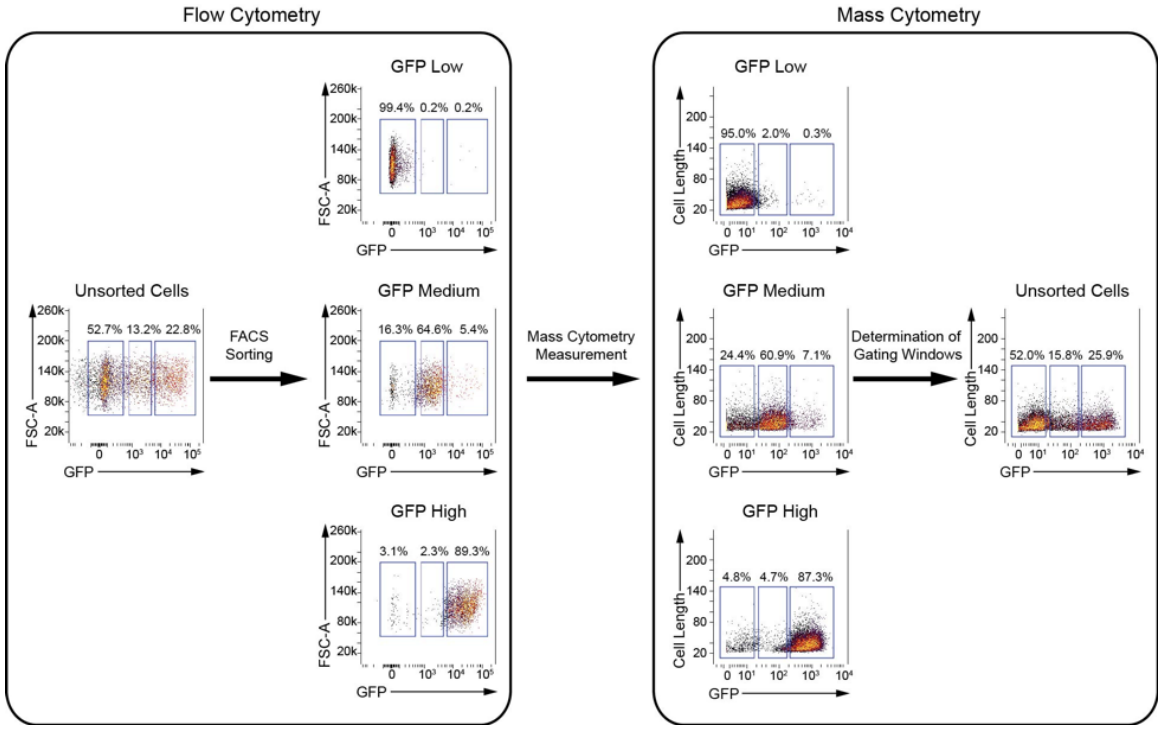


Figure A.5: Comparison of mass cytometry and flow cytometry (FACS). HEK293T cells were transfected with the FLAG-GFP overexpression vector. With flow cytometry, cells were gated into GFP low, medium, and high populations with the gating strategy shown in the left panel. With mass cytometry, each of the three sorted populations was measured independently to determine the gating windows. Unsorted cells were then assessed by the mass cytometry. The maximum difference in population percentage between mass cytometry and flow cytometry was less than 3%.

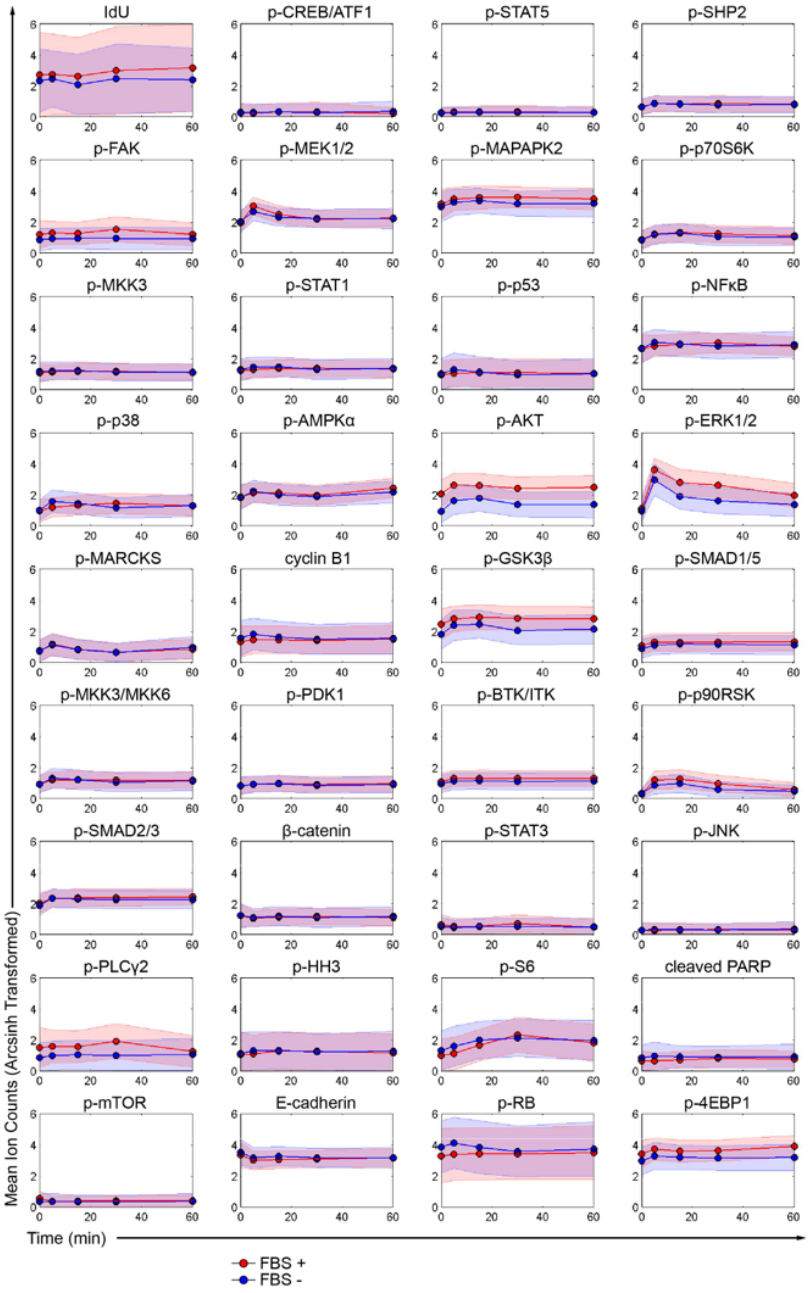


Figure A.6: Comparison of EGF stimulations in starved (FBS is absent) and non-starved (FBS is present) cell culture conditions. HEK293T cells were stimulated with EGF with or without FBS over a 1-h time course. In the non-starved condition basal signaling states of the major MAPK/ERK or AKT pathway components were higher than in starved conditions, but these elevated levels did not affect the signaling responses to the EGF stimulation. Mean value of each sample is shown with circle. Standard deviation is indicated by shaded area.

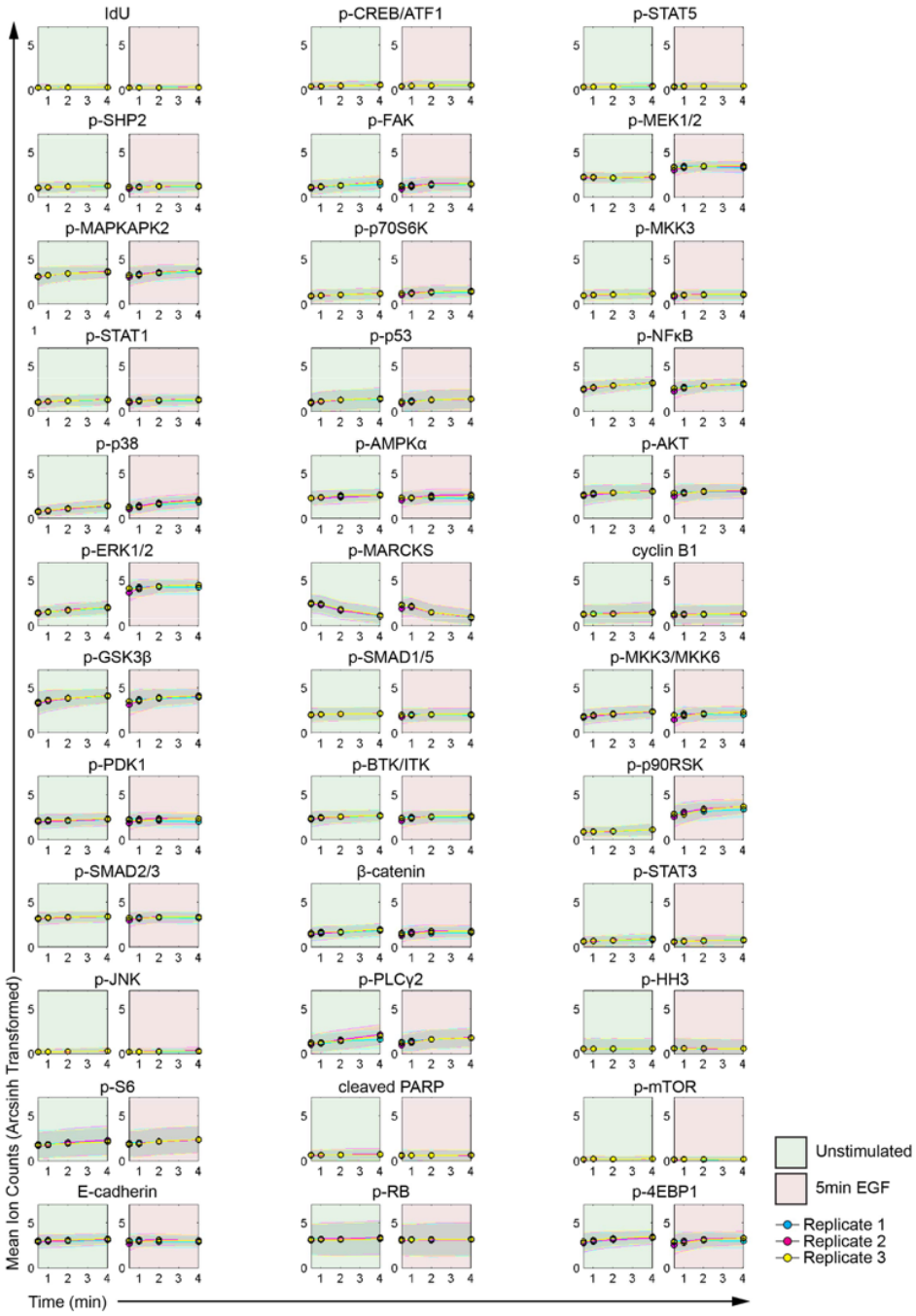


Figure A.7: TrypLE treatment time course. HEK293T cells were treated with TrypLE for 30 s, 1 min, 2 min, or 4 min with or without EGF stimulation for 5 min (time from EGF addition to PFA crosslinking). Within the first 2-min TrypLE treatment (i.e., the time after which we quenched cells in all experiments), only phosphorylation of Ser167/170 on MARCKS varied relatively. Mean value of each sample is shown with circle. Standard deviation is indicated by shaded area.

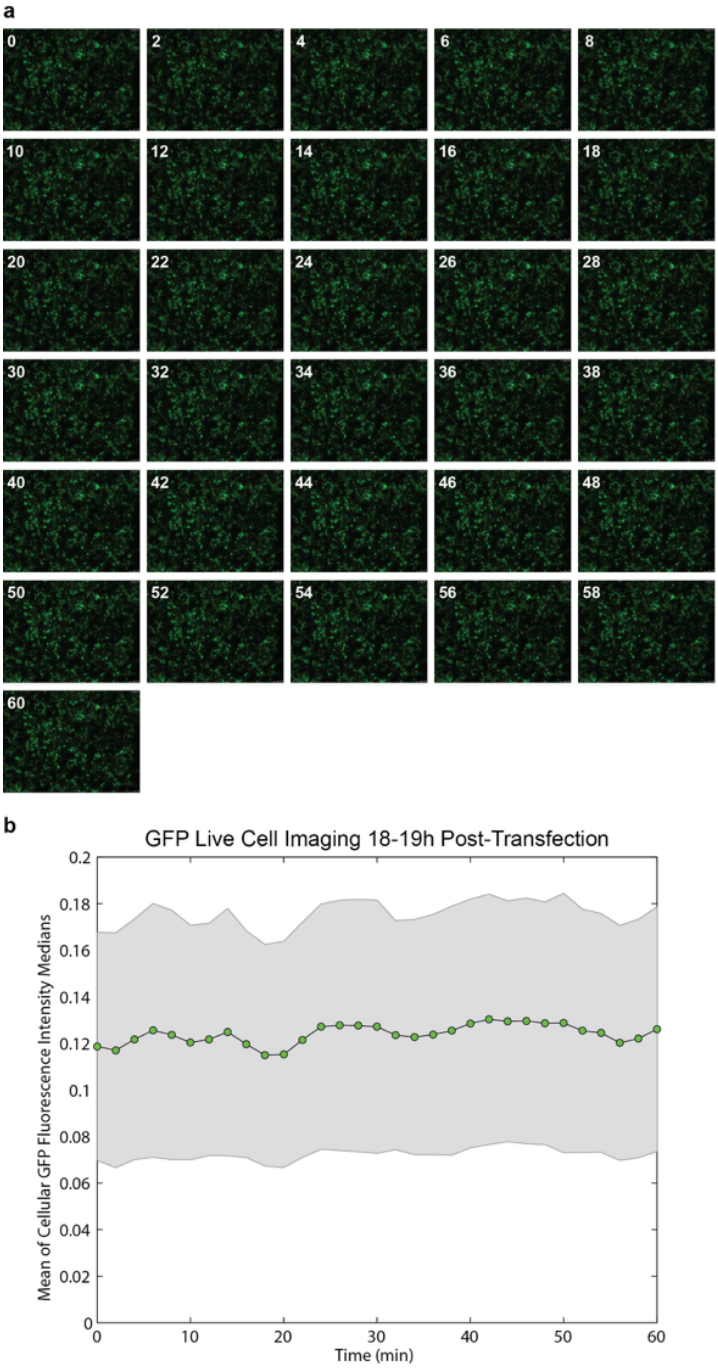


Figure A.8: Live imaging of GFP fluorescence at 18 to 19 h after HEK293T cells were transfected with a FLAG-GFP construct. Quantification of the GFP intensity showed a slight increase of 5.4% over the 1-hr time course. There was a fluctuation in total GFP signal, indicating that the 5.4% increase is most likely attributable to technical variability of the measurement. The analysis of signaling relationships in our study was performed based on a binning strategy on arcsinh transformed GFP ion counts (mass cytometry). Thus, the measured change will not significantly affect the binning over the time course. Standard deviation is indicated by shaded area.

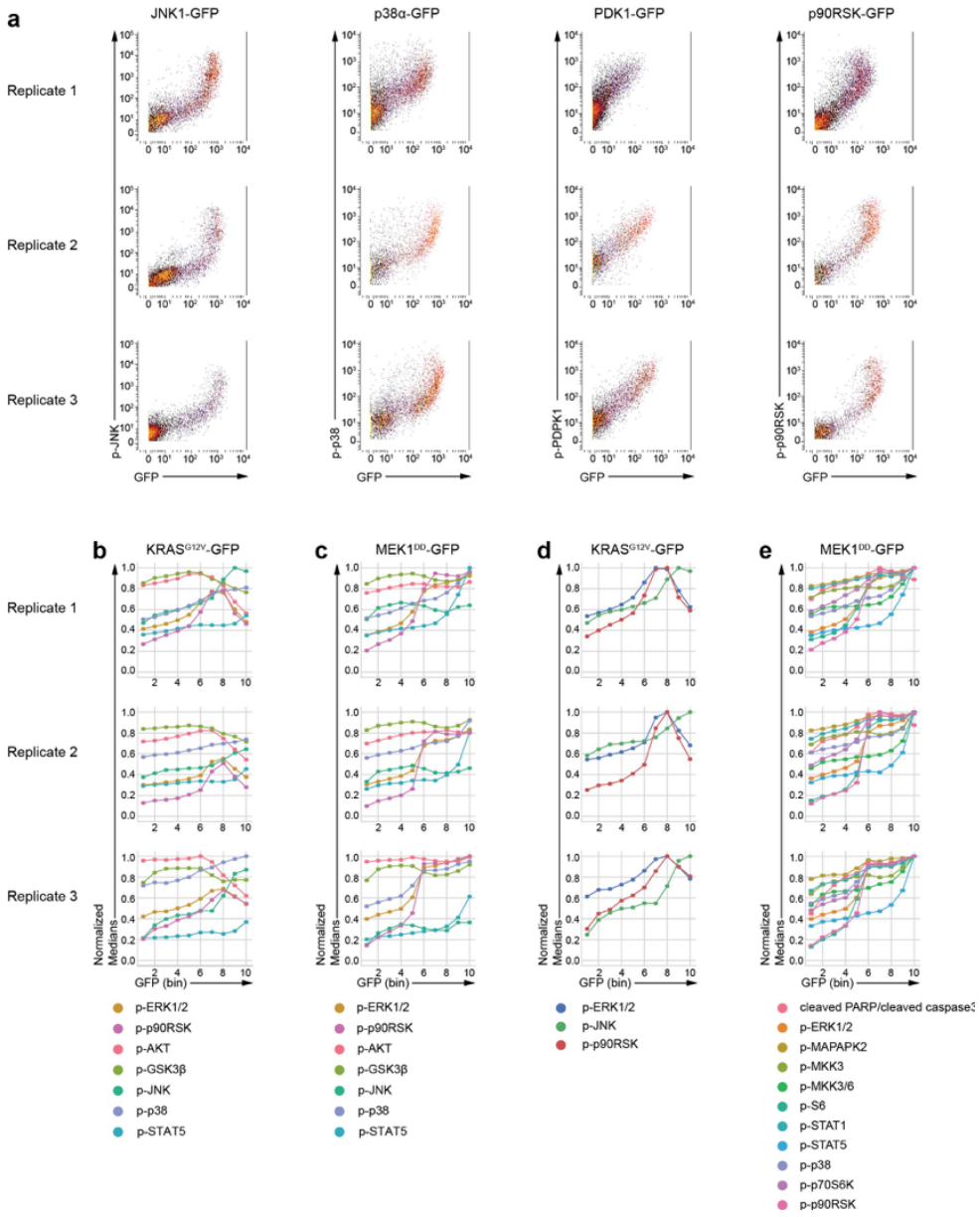


Figure A.9: Abundance-dependent signaling analyses performed in individual experiment replicates are highly reproducible. **(a)** Different batches of HEK293T cells were transfected with JNK1-GFP, P38α-GFP, PDK1-GFP, or p90RSK-GFP constructs, stained, and analyzed by mass cytometry on three different days. Highly consistent signaling responses were observed among the three individual experiment replicates. Panels **(b)** and **(c)** show analyses of representative phosphorylation sites in the MAPK/ERK, AKT, stress pathways, and the STAT5 protein in cells in which **(b)** $KRAS^{G12V}$-GFP and **(c)** $MEK1^{DD}$-GFP was overexpressed. Panels **(d)** and **(e)** show all relationships that passed the BP-R^2 threshold (see Methods for details) for the **(d)** $KRAS^{G12V}$-GFP and **(e)** $MEK1^{DD}$-GFP overexpression experiments.

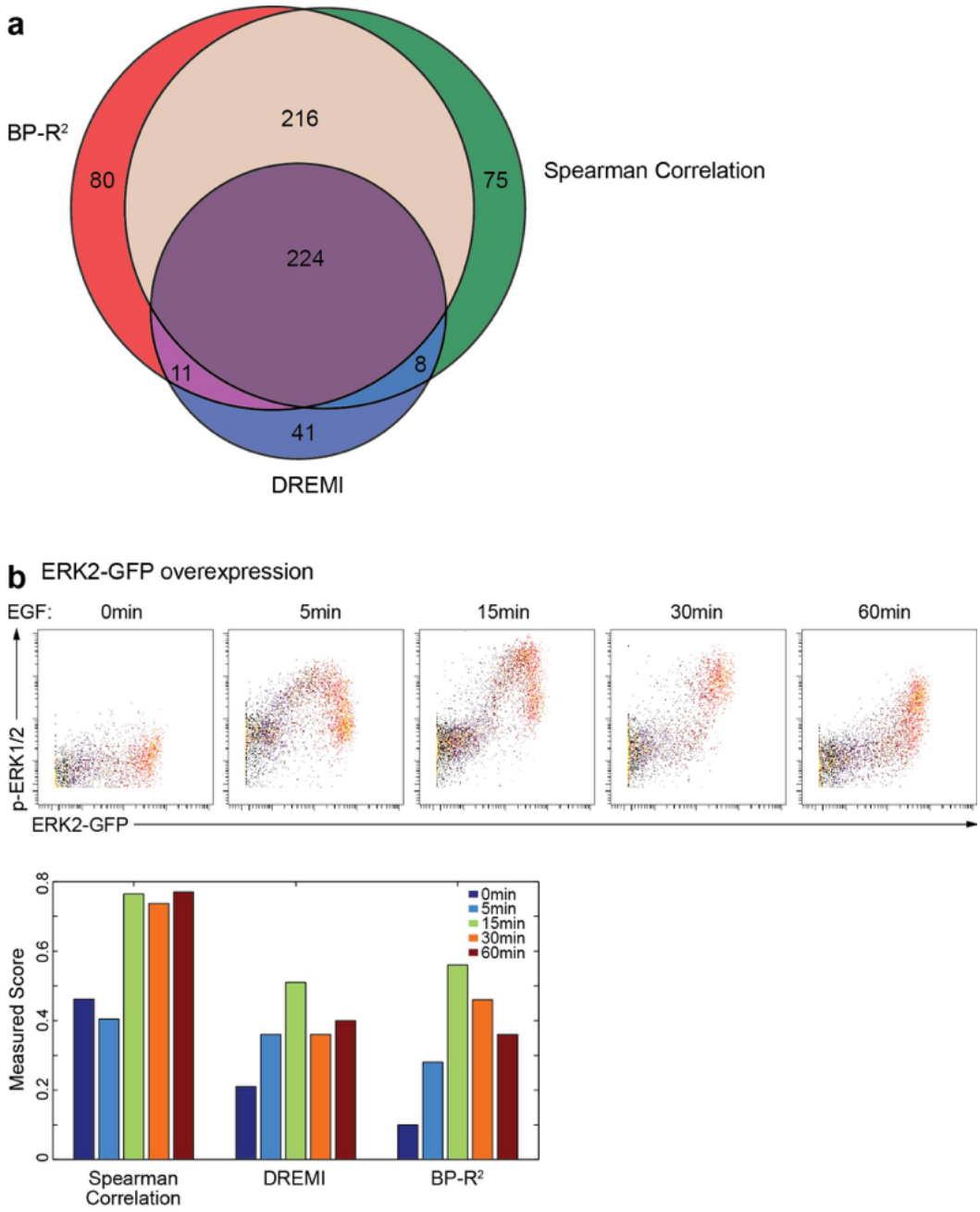


Figure A.10: Benchmark of BP-R^2 against other methods used to identify relationships in mass cytometry data. (a) Venn diagram of strong relationships detected by BP-R^2, Spearman correlation, and DREMI in our dataset using the same cutoff - the 99$^{\text{th}}$ percentile of the BP-R^2 / Spearman correlation / DREMI score in the control groups (FLAG-GFP overexpression and the untransfected cells). BP-R^2 outperforms the other two measures. **(b)** BP-R^2, Spearman correlation, and DREMI measurements of signaling relationship strength between p-ERK1/2 and overexpressed ERK2-GFP. BP-R^2 is suitable for analyzing non-monotonic signaling relationships and outperforms the other two measures in representing actual signaling activation status.

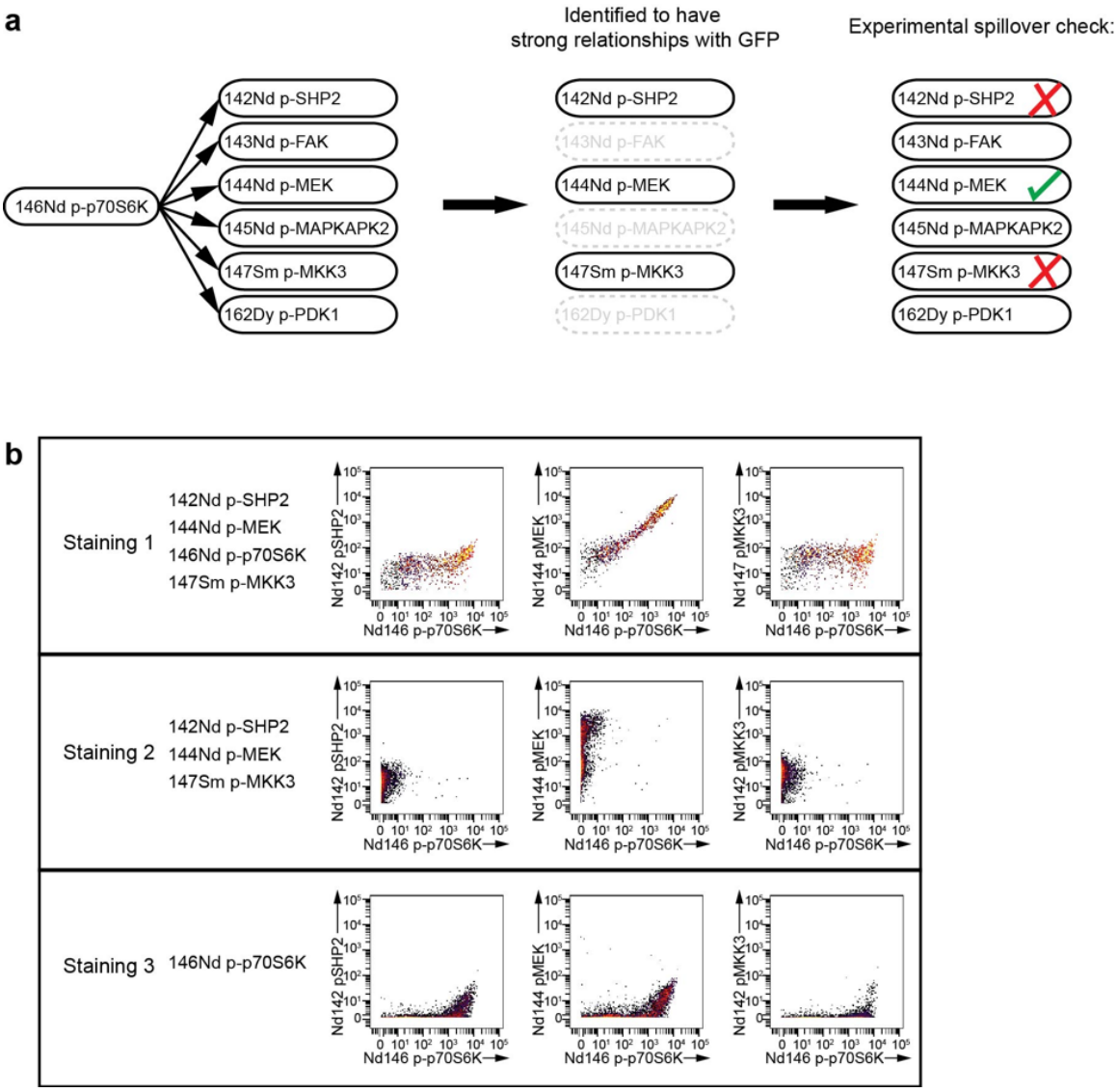


Figure A.11: Analysis of signal spillover among mass channels. (a) Strategy to exclude spillover among mass channels. When strong signaling relationships as determined by BP-R^2 were identified (measured phosphorylation of p70S6K in the p90RSK-GFP overexpression is shown here as a selected example), all other potentially affected channels (details in Methods) were evaluated for spillover that might have led to a high BP-R^2 value. Using an experimental spillover filter **(b)**, spillover-affected relationships were discarded. Here three groups of antibody stains were performed simultaneously: First, all antibodies; second, all antibodies except for the one that potentially causes spillover; third, only the antibody that potentially causes spillover. If spillover induced background was over 10% of the actual ion counts, the channel was discarded from the analysis.

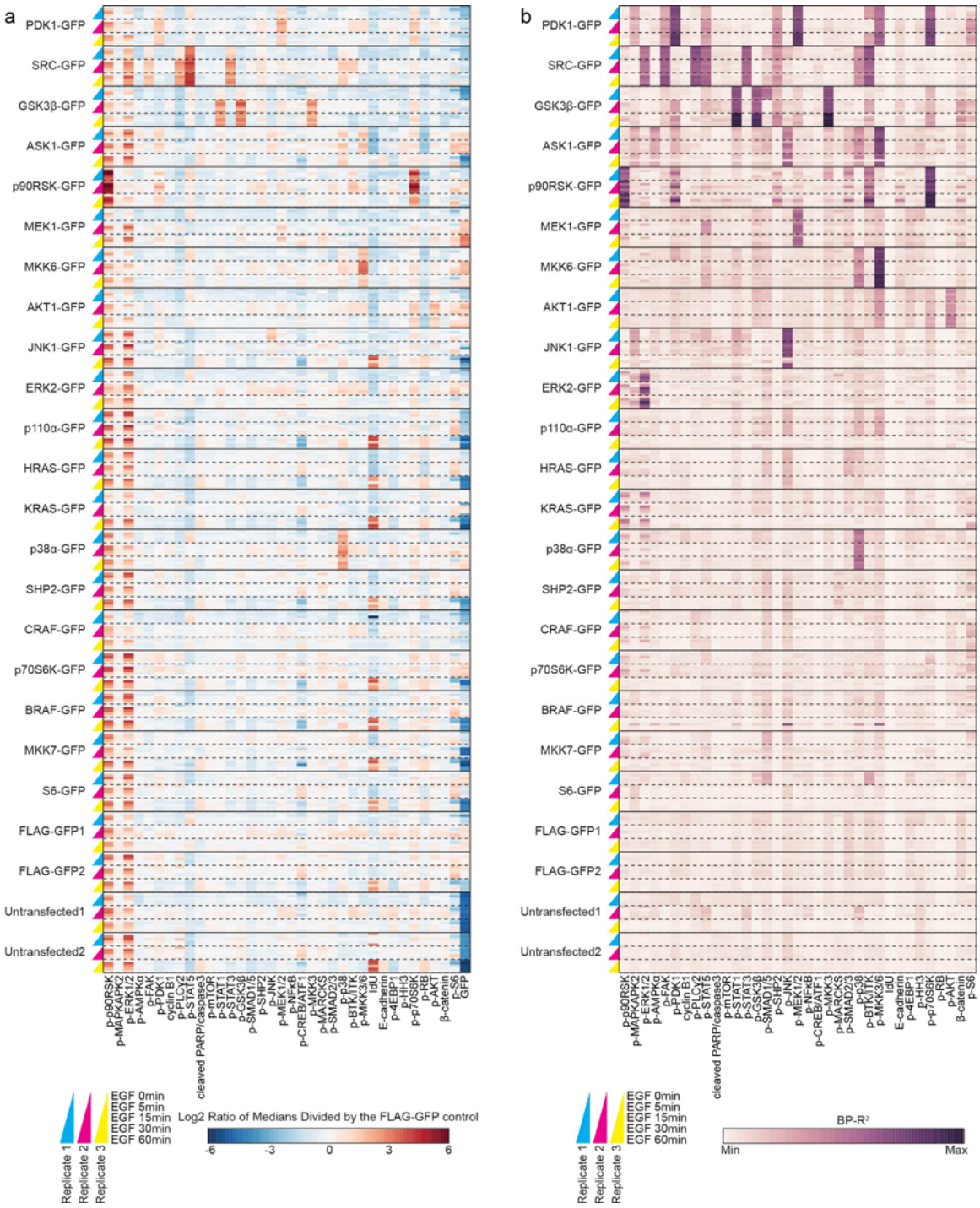


Figure A.12: Median intensities and BP-R^2 analysis for all experimental conditions. (a) Heat map of median intensities of all measured markers at 0, 5, 15, 30, and 60 min post-EGF stimulation in all overexpression conditions (Table 2.1). Data visualized as $\log_2$ of the ratio of the median signals divided by the mean of median signals of the FLAG-GFP controls at time point 0. **(b)** Heat map of BP-R^2 values of all measured markers versus GFP signals at 0, 5, 15, 30, and 60 min post-EGF stimulation in all overexpression conditions.

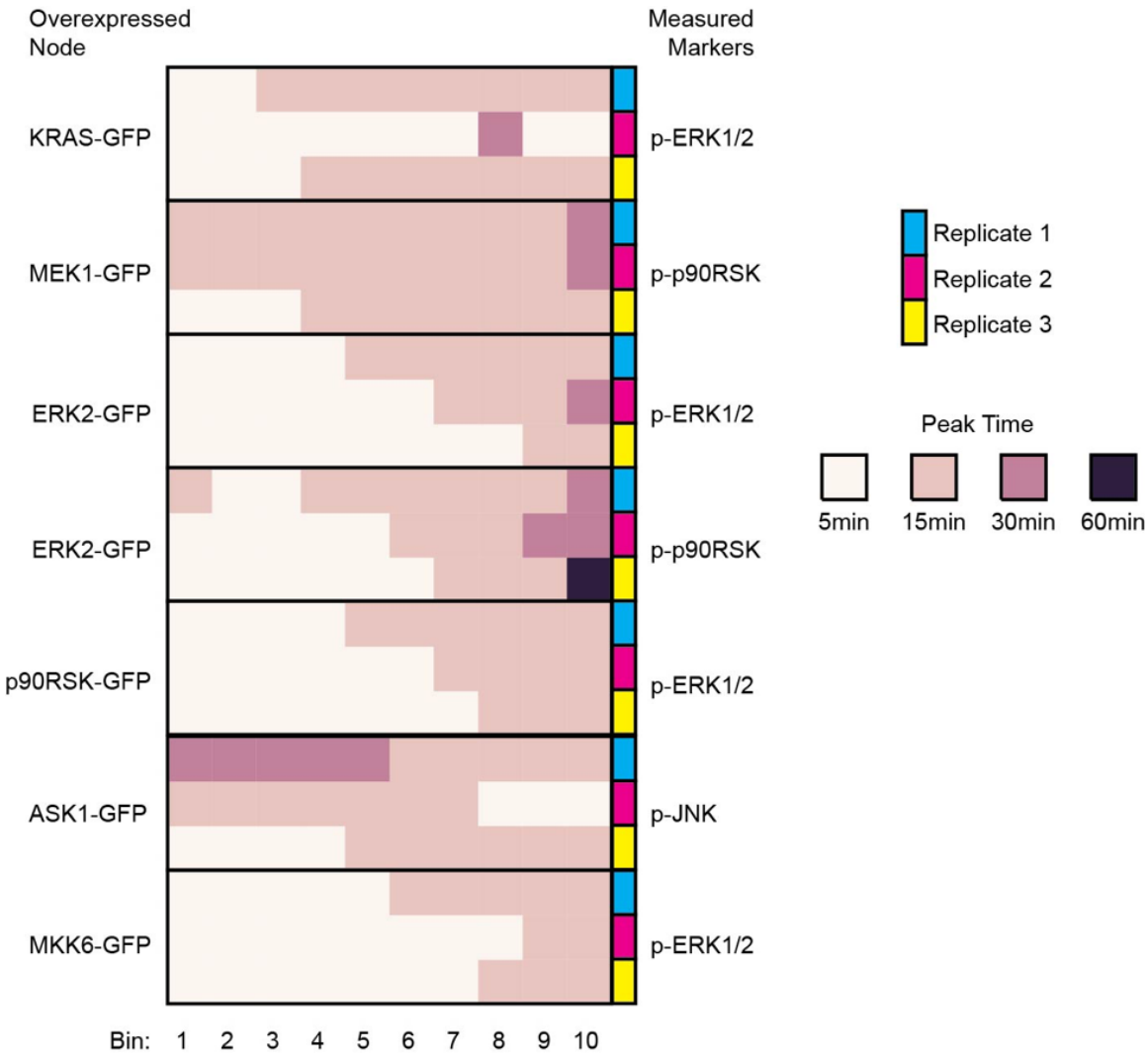


Figure A.13: Strong and robust changes in signaling peak times. Heat map of consistent and robust examples for overexpression-induced phosphorylation site abundance peak time changes after EGF stimulation for each of the three replicates.

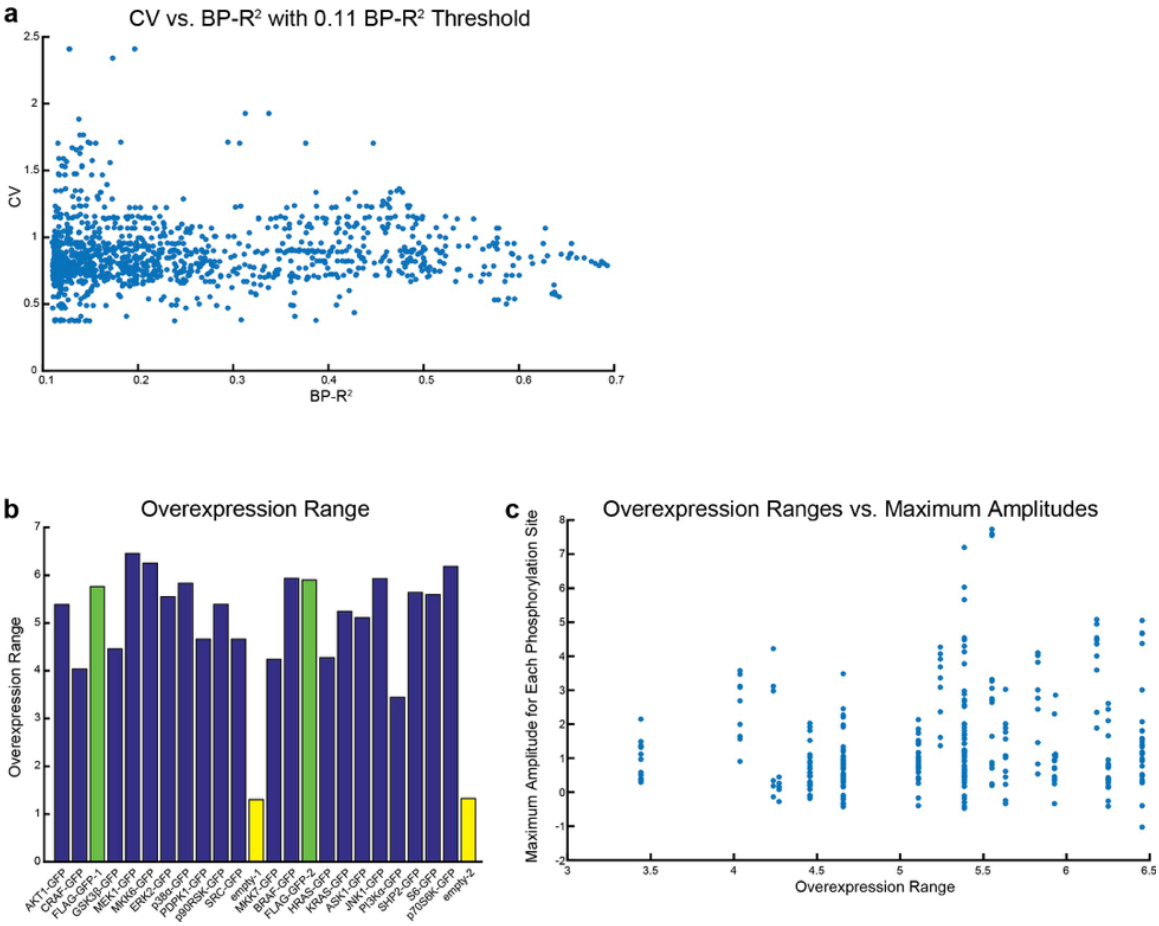


Figure A.14: Post-transcriptional constraint analysis of overexpressed POIs. (a) Coefficient of variation (CV) was computed for each strong signaling relationship that had a BP-R^2 value above 0.11 (i.e., a strong signaling relationship), and CVs were plotted against BP-R^2. No correlation was observed. **(b)** Overexpression ranges (median value of GFP in Bin_{10} minus the median value of GFP in Bin1) calculated for all POIs. **(c)** Maximum amplitudes of phosphorylation sites were independent of the level of overexpression of the POIs.

A.2 Supplementary Tables

All supplementary tables (1-6) may be found in the online methods (see (Lun et al., 2017)).

APPENDIX B

SUPPLEMENTARY MATERIALS: VARIATION IN SINGLE-CELL SIGNALING DYNAMICS IS DETERMINED BY INITIAL CELL STATE

B.1 Introduction

Here, we provide additional details on the 'single-cell ODE modeling' (SCODEM) method presented in the main text. This material includes a description of the experimental and modeling workflow, methods for scaling and subsampling the raw experimental data for use in modeling and, finally, the mathematical model structures used.

B.2 SCODEM

Single-cell ODE modeling is based on the assumption that the majority of variation observed in the single-cell dynamics (trajectories) of a reaction system originates from and, therefore, is explained by the cell-to-cell variation in the initial sate of the system. This assumption entails that possibly stochastic reaction kinetics has a minor effect on dynamics and is the main reason for applying SCODEM to study signal transmission rather than translation. (We note, however, that the general single-cell modeling methodology presented here may readily be adapted to study gene regulatory systems; by using stochastic differential equations, for example.) Thus, to predict the forward-time progression of system state variables of a given system, we assume that knowledge of the variation in initial system states is sufficient.

As an illustration, here we present a single step of the SCODEM approach in the simplest case, namely, where we assume that a valid model structure of the system is known in the form of an appropriate approximation of the system structure. Moreover, we assume that all state variables in the model have been experimentally measured in a number of single cells sufficient for statistically approximating the distribution structure of the underlying

cell population(s). We will subsequently describe how SCODEM may be applied to more realistic cases, as discussed in the main text, where we may not know all state variables that should be included in a model or where we cannot experimentally measure some of the state variables in the model.

B.2.1 SCODEM algorithm

We consider a dynamical biochemical system $\mathbb{S}$. The states of $\mathbb{S}$ are characterized by the vector of state variables $\mathbf{X} \in \mathbb{R}^v$ and the system of ODEs defining $\dot{\mathbf{X}}$ sufficiently approximates $\mathbb{S}$. Given a vector of initial system states $\mathbf{X}_0$, a vector of system parameters $\mathbf{\Theta}$ and a vector of system inputs $\mathbf{u}$, the solution of the ODE system $\dot{\mathbf{X}}(\mathbf{X}_0, \mathbf{\Theta}, \mathbf{u})$ describes the time evolution (dynamics) of the state variables $\mathbf{X}(t)$ representing system $\mathbb{S}$.

Additionally, we consider multiplexed single-cell experimental snapshot data $\mathbf{D}_t$ measured at times $t \in T$ such that the dataset $\{\mathbf{D}_t, t \in T\}$ characterizes the dynamics of the system $\mathbb{S}$. Specifically, an individual snapshot $\mathbf{D}_t$ is an $n \times v$ matrix of v state variables observed in n cells at experimental time t. We assume that the number of cells n in each measurement is sufficient to characterize the structure of the underlying "true" population.

In this simplest case, where the model structure is known, the SCODEM algorithm is concerned with inferring the biochemical rate parameters $\mathbf{\Theta}$ of the ODE system $\dot{\mathbf{X}}(\mathbf{X}_0, \mathbf{\Theta}, \mathbf{u})$. Any single step s of SCODEM is implemented in an optimization routine of the following type:

Algorithm 1 Generic optimization step s of SCODEM

1: **Input:** $\mathbf{\Theta}_s$, $\{\mathbf{D}_t, t \in T\}$, $\mathbf{u}$

2: **Output:** Cost_s

3: **procedure** INTEGRATE SINGLE-CELL INSTANCES OF ODE SYSTEM

4: **for** $k = 1 : n$ **do** ▷ For each cell k measured at time t_0

5: $\hat{\mathbf{X}}_k(t) \leftarrow \int_{t_0} \dot{\mathbf{X}}(\mathbf{D}_{t_0,k}, \mathbf{\Theta}_s, \mathbf{u}_k)dt$ ▷ simulate trajectory $\hat{\mathbf{X}}_k(t)$ of k^{th} cell in $\mathbf{D}_{t_0}$ according to $\mathbf{\Theta}_s$

6: **end for**

7: **end procedure**

8: $\hat{\mathbf{X}}(t) = \bigcup_{k=1}^{n} \hat{\mathbf{X}}_k(t)$ ▷ Combine n simulated single-cell trajectories to define trajectory of empirical distribution

9: **procedure** COMPARE SIMULATED AND EXPERIMENTAL DISTRIBUTION SNAPSHOTS

10: $\text{Cost}_s = \sum_{t \in T} \text{MMD}(\hat{\mathbf{X}}(t), D_t)$ ▷ Sum maximum mean discrepancy values between simulated and measured distributions at measurement times

11: **end procedure**

12: **Output:** Cost_s

B.2.2 Inference of latent cell-to-cell variation

In real applications of SCODEM, we typically do not know the full system structure of $\mathbb{S}$ and/or cannot measure all relevant state variables. Expressed differently, the dimensionality of the model $\mathbf{X}^*$ or measurements $\mathbf{D}^*$ is less than the full system. In either case, the corresponding ODE model $\dot{\mathbf{X}}^*(\mathbf{X}_0^*, \boldsymbol{\Theta}, \mathbf{u})$ that approximates $\mathbb{S}$ can naturally not account for some source(s) of variation in initial cell states that may influence the system dynamics, and the inference procedure may fail to reproduce the variation in system dynamics characterized by snapshot measurements. If the unknown/unmeasurable model variables change on a time scale qualitatively separable from the dynamics being studied, however, it is possible to infer certain characteristics of the cell-to-cell variation in the unknown or unmeasured variables. This inference is made possible by allowing some model parameters to vary across cells, and by determining the values of these parameters by analytical solution of the steady-state equations using experimental measurements.

Specifically, we take a parameter subset $\boldsymbol{\Phi} \subseteq \boldsymbol{\Theta}$ and allow these parameters ($\boldsymbol{\Phi}_k$) to vary across each cell k in a population, while keeping the remaining parameters the same for all cells. For a given estimation of population parameters $\boldsymbol{\Theta}$, the individual cell values of $\boldsymbol{\Phi}_k$ are algebraically determined by steady-state solution of the ODE system using the current estimate of population parameters $\boldsymbol{\Theta}$ and steady-state measurements $\mathbf{D}^*_{t_0,k}$ of cell k.

We use an example from our model to illustrate the procedure. We begin with the

following differential equation for ERK activation:

$$\dot{ppERK} = k_5 * ERK^{g_3} * ppMEK^{g_4} - k_6 * ppERK^{h_2} \tag{7}$$

Because the kinetic order parameters g_3, g_4 and h_2 are directly related to physical properties of system components explicitly represented in the model as state variables, we assume they do not change across the cell population (thus, $\{g_3, g_4, h_2\} \in \mathbf{\Theta}$). If our reaction scheme for ERK activation were complete, we would expect that the rate constant parameters k_5 and k_6 to be the same across the population. Our reaction scheme, however, is an approximation of the true reaction structure and does not include all known and unknown contributors to the reaction; in particular, mono-phosphorylated ERK in the forward reaction and ERK phosphatase in the reverse reaction. For our given experimental conditions, we begin with the assumption that our model explicitly captures the primary state variables that change their state on the same time scale. For instance, we assume that mono-phosphorylated ERK may rapidly equilibrate and the phosphatase is constitutively active. If our assumption is incorrect, our model will not be able to fit the data and we know our model structure must be updated. By contrast, if our assumption is a valid approximation for our experimental conditions, then we may use the parameters k_5 and k_6 to capture the sources of latent cell-to-cell variation. Thus k_5 and k_6 are included in the single-cell parameters set $\mathbf{\Phi}_k$. The single-cell parameter values $k_{5,k}$, taken across all cells, capture variation in mono-phosphorylated ERK and any unknown ERK activators, while $k_{6,k}$ capture the variation in ERK phosphatase and any unknown ERK inactivators. Given some guess of $\mathbf{\Theta}$, which may be set based on biological knowledge or determined by the optimization algorithm, we may use the model structure from equation (7) and our measurement of cell k at steady state ($\mathbf{D}^*_{t_0,k}$) to infer some properties of single-cell parameter values $k_{5,k}$ and $k_{6,k}$ by solving the

steady state equation:

$$\text{Steady state} \Rightarrow 0 = \dot{ppERK}_k = k_{5,k} * ERK_{t_0,k}^{g_3} * ppMEK_{t_0,k}^{g_4} - k_{6,k} * ppERK_{t_0,k}^{h_2} \quad (8)$$

$$\Rightarrow \frac{k_{6,k}}{k_{5,k}} = \frac{ERK_{t_0,k}^{g_3} * ppMEK_{t_0,k}^{g_4}}{ppERK_{t_0,k}^{h_2}} \quad (9)$$

The parameters $k_{5,k}$ and $k_{6,k}$ are non-identifiable at the steady state due to the model structure, and we only identify the ratio. If one parameter, say $k_{5,k}$, is fixed, however, the other ($k_{6,k}$) is fully determined by the model structure, the current population-level parameter values $\mathbf{\Theta}$, and the constraint that the system must be at a steady state with the current (measured) combination of state variables. The fixed parameter $k_{5,k}$, while non-identifiable at the steady state, serves to scale the speed of the forward and reverse reactions and can therefore be inferred from the time course measurements of the system dynamics. If we assume the free parameter $k_{6,k}$ to be responsible for capturing all the cell-to-cell variation in the ERK activation/inactivation reaction, then parameter $k_{5,k}$ can be added back to the population-level parameter set ($k_{5,k} \rightarrow k_5 \in \mathbf{\Theta}$). The result is that, first, the number of unknown kinetic parameters for the reaction system describing the change in active ERK is decoupled from the number of cells and, secondly, that the union of single-cell parameter values $k_{6,k}$ for all cells k in the sample defines a distribution k_6 that captures the unmeasured/unknown sources of cell-to-cell variation in the reaction scheme. Given additional measurement features, we may be able to determine further contributors to the reactions of ERK activation by looking at single-cell correlations between these features and the inferred distribution of k_6 ($k_{6,k}$ across all n cells).

B.3 Data processing for use in modeling

SCODEM uses single-cell measurement values to instantiate ODE model instances. This approach is advantageous because it is simple; however, it requires attention to errors in individual cell measurements and, in the case of mass cytometry, rescaling and normalization of measurement channels for use in modeling.

B.3.1 Data exclusion criteria and resampling

Antibody labeling and cell staining was optimized to minimize the number of cell events with zero or low ion counts in measurement channels used for modeling (e.g., total ERK). Furthermore, measured cell events with fewer than 5 ion counts in modeled channels were excluded. The choice of five counts was made as a trade-off between the number of discarded events, which increases as the threshold increases, and the uncertainty of single-cell measurements, which is a decreasing function of ion counts. Such thresholding may introduce bias across samples if, for example, the active form of a protein is on average low in abundance at one time point, but increases in abundance upon stimulation. In this case, thresholding may remove many "low" cells from the first time point that ultimately should represent some of the cells at later time point. To address this issue, we used a stochastic model of mass cytometry measurement based on models of inductively coupled plasma mass spectrometry (ICP-MS)(Ulianov et al., 2015). Specifically, we modeled the number of measured ion counts X as a Gamma-Poisson mixture, which can be represented by a negative binomial distribution $X \sim NB(r, p)$, where r represents the number of metal tags in the cell before measurement and p is the probability of a given metal tag successfully reaching the detector. This probability is a function of the instrument; in our case, it is $p = 1/5000$. To avoid probability mass at zero, we first shifted all distributions by 10 ion counts before resampling.

In ERK overexpression experiments, the number of ion counts measured could exceed the linear response range of the mass cytometer. As nonlinearities in measurement violate the optimization objective function as implemented, we excluded cell events with ion counts greater than 10,000 before data resampling and modeling.

B.3.2 Data scaling

Mass cytometry provides relative values of protein abundance. Absolute values depend on factors such as antibody labeling efficiency, antibody staining and detection sensitivity. While relative differences between measured protein abundances can often be subsumed into reaction parameters in a model, we rescaled measurements to physiological values.

Average protein abundance values in HEK293T cells were obtained from the MaxQB database (`http://maxqb.biochem.mpg.de/mxdb/`)(Geiger et al., 2012). The average of iBAQ abundance values across the three experimental replicates was used to determine average abundance (Supplementary Table B.5). In cases where an antibody detected multiple isoforms of a protein (e.g., ERK1 and ERK2) the average abundances of each isoform were summed. Abundance values were scaled to concentrations using the average HEK cell volume(Boss et al., 2013) of 1996 μm^3. These values were used to estimate the average concentration $\bar{\mathbf{x}}_j$ of each protein j in a cell. Protein measurements were then linearly scaled as follows:

Given a protein j, its associated average concentration $\bar{\mathbf{x}}_j$ and its measured steady state distribution $D_{\mathbf{x}_{j|ss}}$, a scaling factor z_j was calculated as

$$z_j = \frac{\bar{\mathbf{x}}_j}{\mathrm{E}\big[D_{\mathbf{x}_{j|ss}}\big]}. \tag{10}$$

This implementation represents a linear scaling factor of the experimentally measured steady-state distribution of protein j such that the mean of the steady-state distribution $\mathrm{E}[D_{\mathbf{x}_{j|ss}}]$ is equal to the population average $\bar{\mathbf{x}}_j$ estimated from the quantitative abundance measurements in MaxQB. For all experimental measurements, e.g., after a perturbation, each total protein j was scaled using the corresponding z_j.

Steady-state levels of phosphoproteins clearly cannot not be greater than those of the corresponding total protein pools. Additionally, if steady-state phospho-protein levels are too high relative to total protein, the relative increase in phosphorylation levels in the model would be capped due to a lack of unphosphorylated protein. To avoid these issues, phospho-protein distributions were scaled such that the average steady-state value of a phospho-protein was a fractional value of the total protein level. The additional scaling factors for active MEK, ERK and p90RSK were 0.08, 0.08 and 0.05, respectively. These values were chosen to reduce the number of cells in violation of the "active cannot be greater than total" constraint at steady state. In the special case that individual cells violated this constraint, these cells were excluded from the sample for analysis.

B.4 Model of EGF signaling in the MAPK/ERK cascade

In this section, we describe the mathematical formulation of our model of EGF signaling in the MAPK/ERK cascade.

B.4.1 State variables

The model uses eight state variables to describe changes in ERK pathway signaling components.

Table B.1: Model variables.

Variable	Description
I	Input (EGF)
pRAF	Active RAF
MEK	Inactive MEK
ppMEK	Active MEK
ERK	Inactive ERK
ppERK	Active ERK
P90	Inactive p90RSK
pP90	Active p90RSK

Active and total MEK, ERK and p90RSK were measured. After scaling (section SB.3.2), inactive forms were calculated as Inactive = Total − Active. As cells were grown in a monolayer and EGF was rapidly mixed with medium, the input I was assumed to be the same for all cells. The initial value of pRAF was also assumed to be the same for all cells. This choice was made to reduce prior assumptions on the distribution of active RAF. The variation in initial pRAF, as well as of other components upstream of MEK that were not explicitly in the model, was captured by the single-cell parameters k_d and k_4, which were obtained from steady-state measurements (as in section B.2.2).

B.4.2 Kinetic parameters

Table B.2: Model parameters.

Parameter	Description
k_f	Activation rate constant of pRAF by Input
k_d	Degradation rate constant of pRAF signal
g_D	Kinetic order of pRAF activation from Input
k_{FB}	Hill function inflection point of negative feedback from ppERK
f_n	Hill function order of negative feedback from ppERK
k_3	Activation rate constant of ppMEK by pRAF
k_4	Inactivation rate constant of ppMEK
g_1	Kinetic order of ppMEK activation from MEK
g_2	Kinetic order of ppMEK activation from pRAF
h_1	Kinetic order of ppMEK inactivation from ppMEK
k_5	Activation rate constant of ppERK by ppMEK
k_6	Inactivation rate constant of ppERK
g_3	Hill function order of ppERK activation from ERK
g_4	Kinetic order of ppERK activation from ppMEK
h_2	Kinetic order of ppERK inactivation from ppERK
k_7	Activation rate constant of pP90 by ppERK
k_8	Inactivation rate constant of pP90
g_5	Kinetic order of pP90 activation from P90
g_6	Kinetic order of pP90 activation from ppERK
h_3	Kinetic order of pP90 inactivation from pP90

The rate constants k_d, k_4, k_6 and k_8 were in $\mathbf{\Phi_k}$ and computed from the steady-state equations as shown in section B.2.2. All other parameters were in $\mathbf{\Theta}$ and, therefore, equal across all cells and used as decision variables in the optimization algorithms.

B.4.3 Inputs

The MAPK/ERK pathway components considered in our model were in a pseudo-steady state at time scale of our experiments (less than one hour). Thus, our model must also be at steady state before simulated addition of EGF. We arbitrarily chose the input value of

Table B.3: Model inputs.

Input	Description
I	Input (EGF)
I_{ss}	Input at steady state (before EGF addition)
I_{final}	Minimal EGF level after EGF addition
τ	Time delay between addition of EGF and initial pRAF signaling

$I = I_{ss} = 1$ as the pre-stimulation steady state input I to our model. Addition of EGF was simulated as an instantaneous increase in the input to $I = 10$. We used a time delay τ to represent the delay between experimental addition of EGF to the medium and the time when the signal reached the MAPK/ERK cascade. In other words, τ represents the time it takes for "signal" to pass be transmitted by reactions, such as receptor-ligand binding, receptor activation, etc., and reach RAF activation. Although the exact value of τ for each cell is undoubtedly variable across the population, as not all cells will encounter the EGF signal at the same moment, we assumed each cell to have the same delay (τ) to maintain the entirely deterministic nature of our model. We used $\tau = 2$ minutes based on live cell studies of ERK activation(Ryu et al., 2015). Thus, in the model, if time $t < \tau$, then Input $I = I_{ss} = 1$. Once time $t = \tau$, then input is reset to represent EGF addition and Input $I = 10$. At all times otherwise, input I is a dependent variable and determined by solution of the ODE system. Finally, we found steady-state signaling after addition of EGF was marginally increased compared to pre-stimulation conditions. Thus, for values of $t > \tau$ we used the value $I_{ss} = I_{final} = 1.2$ (see model equations, below) to limit the depletion of input signal. Expressed differently, before the addition of EGF, the minimum signal can be 1, but after addition of EGF the minimum signaling can be 1.2. This choice was an approximation based on data and fixed before parameter optimization.

B.4.4 Equations

$$
\begin{aligned}
\dot{I} &= -k_f * (I - I_{ss}) \\
p\dot{RAF} &= k_f * I - k_d * pRAF^{g_D} * \frac{ppERK^{fn}}{{k_{FB}}^{fn} + ppERK^{fn}} \\
pp\dot{MEK} &= k_3 * MEK^{g_1} * pRAF^{g_2} - k_4 * ppMEK^{h_1} \\
\dot{MEK} &= -pp\dot{MEK} \\
pp\dot{ERK} &= k_5 * ERK^{g_3} * ppMEK^{g_4} - k_6 * ppERK^{h_2} \\
\dot{ERK} &= -pp\dot{ERK} \\
p\dot{P}90 &= k_7 * P90^{g_5} * ppERK^{g_6} - k_8 * pP90^{h_3} \\
\dot{P}90 &= -p\dot{P}90
\end{aligned}
$$

B.5 Modeling ERK overexpression

B.5.1 Addition of saturating functions

The only saturating reaction in our original model was the negative feedback from active ERK to active RAF, which degrades the input signal. The choice to use saturation in negative feedback was due to strong negative feedback effects leading to mild oscillations. Otherwise, however, for the range of expression we observed in "normal" cells, saturating reactions were not necessary to achieve good fits to experimental data characterizing the signaling dynamics. As cancers are often caused by protein expression beyond "normal" ranges, however, it became necessary to examine MAPK/ERK pathway signaling outside of these ranges.

To allow this extrapolation, we first modeled ERK activation as a saturating (ultrasensitive) Hill function of ERK:

$$
pp\dot{ERK} = \frac{k_5 * ERK^{g_3}}{ERK^{g_3} + {k_{ERK}}^{g_3}} * ppMEK^{g_4} - k_6 * ppERK^{h_2},
$$

which added the kinetic parameter k_{ERK} that determines the inflection point of the Hill function. Indeed, this inflection point lies beyond the maximal expression range of total

ERK in "normal" cells.

To reasonably model reaction kinetics during overexpression, which should eventually saturate, we also allowed ERK expression levels to modify kinetic parameters. This approach is justified by the view of a power function as a local approximation to the "true" kinetic function at an operating point. The validity of such an approximation depends on the function and the operational range of the system. For example, based on our results, all "normal" cells within the context of our system and conditions seem to operate within a range where the reaction kinetics is well approximated by the originally chosen, non-saturating power functions. To represent changes in the kinetics related to ERK expression beyond the ranges of "normal" cells, we used the following functions:

$$f_n^* = f_n * \max\left(\frac{ERK_{total}^{n_{fn}}}{ERK_{total}^{n_{fn}} + k_{fn}^{n_{fn}}}, m_{fn}\right) \quad \text{Reduction in negative feedback strength}$$

$$h_2^* = h_2 * \max\left(\frac{ERK_{total}^{n_{h2}}}{ERK_{total}^{n_{h2}} + k_{h2}^{n_{h2}}}, m_{h2}\right) \quad \text{Saturation of inactivating reaction (i.e., phosphatase)}$$

$$k_3^* = k_3 * \max\left(\frac{ERK_{total}^{n_{k3}}}{ERK_{total}^{n_{k3}} + k_{k3}^{n_{k3}}}, m_{k3}\right) \quad \text{Slow activation of MERK}$$

$$k_5^* = k_5 * \max\left(\frac{ERK_{total}^{n_{k5}}}{ERK_{total}^{n_{k5}} + k_{k5}^{n_{k5}}}, m_{k5}\right) \quad \text{Slow activation of ERK}$$

$$k_7^* = k_7 * \max\left(\frac{ERK_{total}^{n_{k7}}}{ERK_{total}^{n_{k7}} + k_{k7}^{n_{k7}}}, m_{k7}\right) \quad \text{Slow activation of p90RSK}$$

Each function modifies the parameter in question as a function of total ERK where the power n_{xx} reflects the steepness and k_{xx} the inflection point of the Hill function. All values of n_{xx} where less than zero, thus the function values ranged from the original parameter value to zero. As reaction velocities should not go to zero, we used the final parameter m_{xx} as the floor of the parameter modification, so that the modified parameters could take values from their original value to m_{xx}, with the actual value a function of ERK.

B.5.2 Simulating ERK overexpression

Cells used in simulating ERK overexpression were subsampled from experimental measurements of control (not ERK2-GFP transfected, not EGF stimulated) conditions. To simulate ERK overexpression resulting from transient transfection, the relative increase in total ERK (ERK_{total}) associated with ERK2-GFP ($ERK2\text{-}GFP$) expression was modeled as a censored gamma distribution with shape parameter $A = 0.154$, scale parameter $B = 86.9$ and lower and upper censoring parameters $l = 0.271$ and $u = 61.4$, respectively. The total ERK

in each cell for the overexpression $ERK_{total_{overexpression}}$ was calculated as:

$$ERK2\text{-}GFP = \min\Big(u, \max\big(l, \Gamma(A, B)\big)\Big),$$
$$ERK_{total_{overexpression}} = ERK_{total} \cdot (1 + ERK2\text{-}GFP)$$

Rate constant parameters were determined from the steady state after the simulated addition of ERK. This choice was made to maintain the steady state and did not qualitatively change the parameter distribution.

B.6 Modeling MEK inhibition

B.6.1 Initial model of CI-1040

The benzhydroxamate compound CI-1040 inhibits MEK non-competitively and thereby reduces the enzymatic v_{max} (maximum velocity) of ERK activation by MEK(Sebolt-Leopold et al., 1999). Originally, we modeled this effect by reducing the rate constant k_5 of ERK activation by MEK. As described in the main text, a simulated dose-response experiment, where the response was defined as a fold-increase of active ERK, revealed a range of CI-1040 sensitivity within the cell populations (Supplementary Figure B.3).

B.6.2 Reconciled model of CI-1040

Experimental treatment of cells with CI-1040 confirmed a range of inhibitor sensitivities within the cell population, but also revealed EGF signaling dynamics at a slower time scale than our initial representation of CI-1040 predicted.

To study the consequences of CI-1040, we fit a model of CI-1040 to a time course of EGF signaling in cells pretreated for 30 minutes with CI-1040 (see Methods). The resulting model of CI-1040 was based on reduction of not only k_5 (ERK activation), but also k_6 (ERK inactivation), k_3 (MEK activation) and k_4 (MEK inactivation). Thus, the effect of CI-1040 was modeled to slow down the activation and inactivation dynamics of both MEK and ERK. While our model successfully captured the delay in ERK activation in the data, it appeared to overestimate the slowing of ERK inactivation in many cases (Supplementary

Figure B.4). This overestimation is likely a consequence of the simplified one-step rather than two-step reaction scheme of ERK activation and CI-1040 inhibition, and one might expect that a more detailed model of the mechanism would be able to fit the data better. Such a model, however, would contain additional parameters and a number of variables not directly measurable in our experiments, such as singly phosphorylated ERK. Importantly, our specific conclusion that initial MEK activity is most predictive of CI-1040 sensitivity depends on the model of CI-1040, and it is conceivable that a different model of CI-1040 could lead to a different conclusion. We did not pursue this aspect of the model in more detail because this portion of our study was meant to represent a proof of principle that single-cell ODE models may be used to analyze and explain differences in response across a population, rather than a detailed model of CI-1040 action. As such, our primary conclusion, that analysis of single-cell ODE models can reveal cooperative combinations of inhibitors, still holds.

B.7 Supplementary Figures and Tables

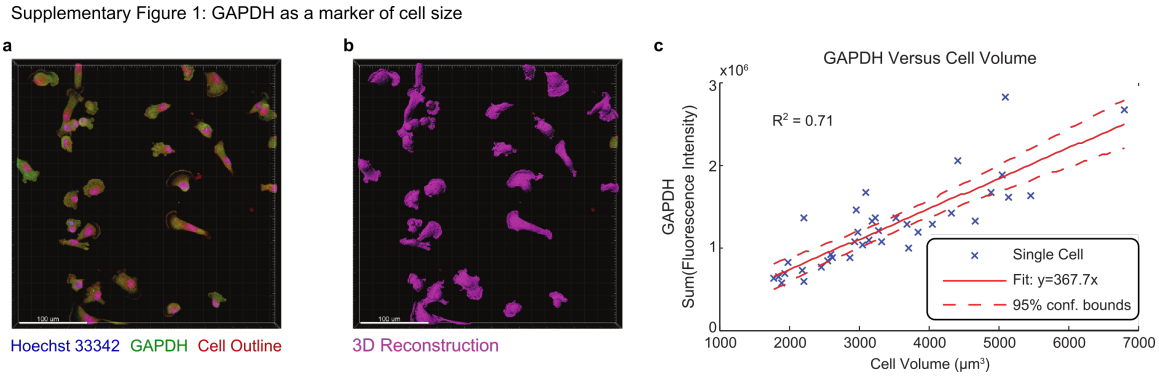


Figure B.1: Correlation between GAPDH and cell volume. (**a**) Confocal images of cells stained for GAPDH (green). Nuclear staining by hoechst 33342 (blue) and Alexa Fluor 647 carboxylic acid succinimidyl ester to determine cell outline (red). (**b**) Three-dimensional reconstruction of cells based on cell outline staining. (**c**) Regression of GAPDH versus reconstructed cell volume.

Supplementary Figure S2: Total protein (including GAPDH) is constant during time course of the experiment

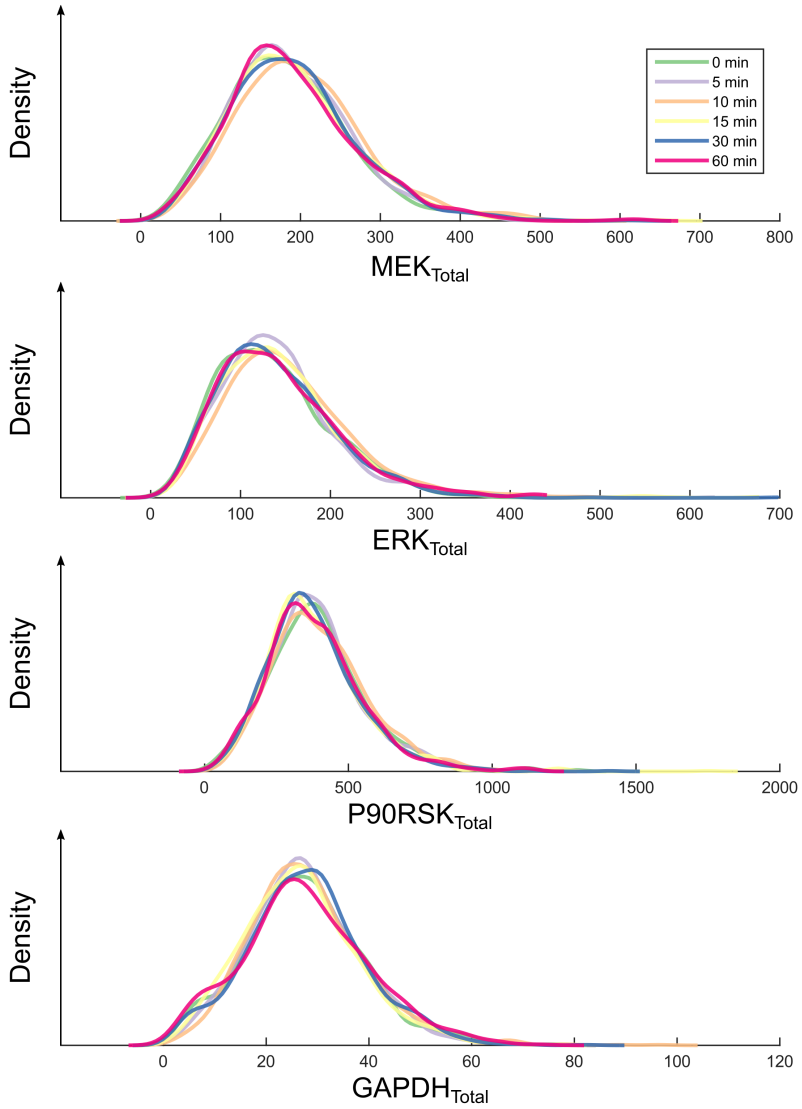


Figure B.2: Total protein abundance does not change during experiment. Snapshots of total protein distributions during a 60 minute time course of EGF stimulation. Units: scaled concentrations.

Supplementary Figure S3: Dose response of original CI-1040 model reveals EGF responsive cells under CI-1040 treatment

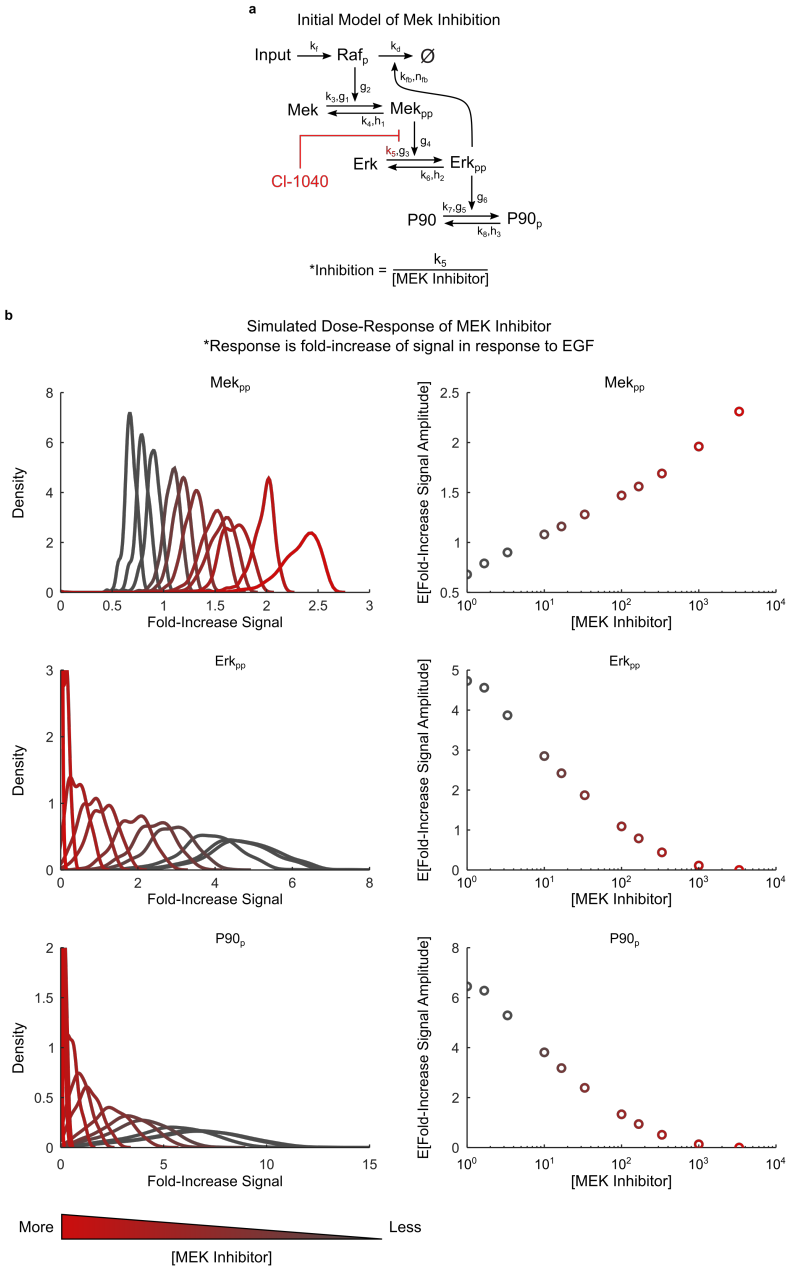


Figure B.3: Simulated dose-response experiment using initial model of CI-1040 MEK inhibitor. Inhibitor acts on parameters in red. (**a**) Model diagram of CI-1040 inhibition. (**b**) Dose response results. Response was defined as fold-increase of signal after EGF simulated stimulation. Color represents amount of inhibitor (red: high; gray: low). (Left) Response distribution of cell population for different levels of simulated CI-1040 inhibition. (Right) Average of response distribution (y-axis) versus level of CI-1040 treatment (x-axis). Model units: scaled concentrations.

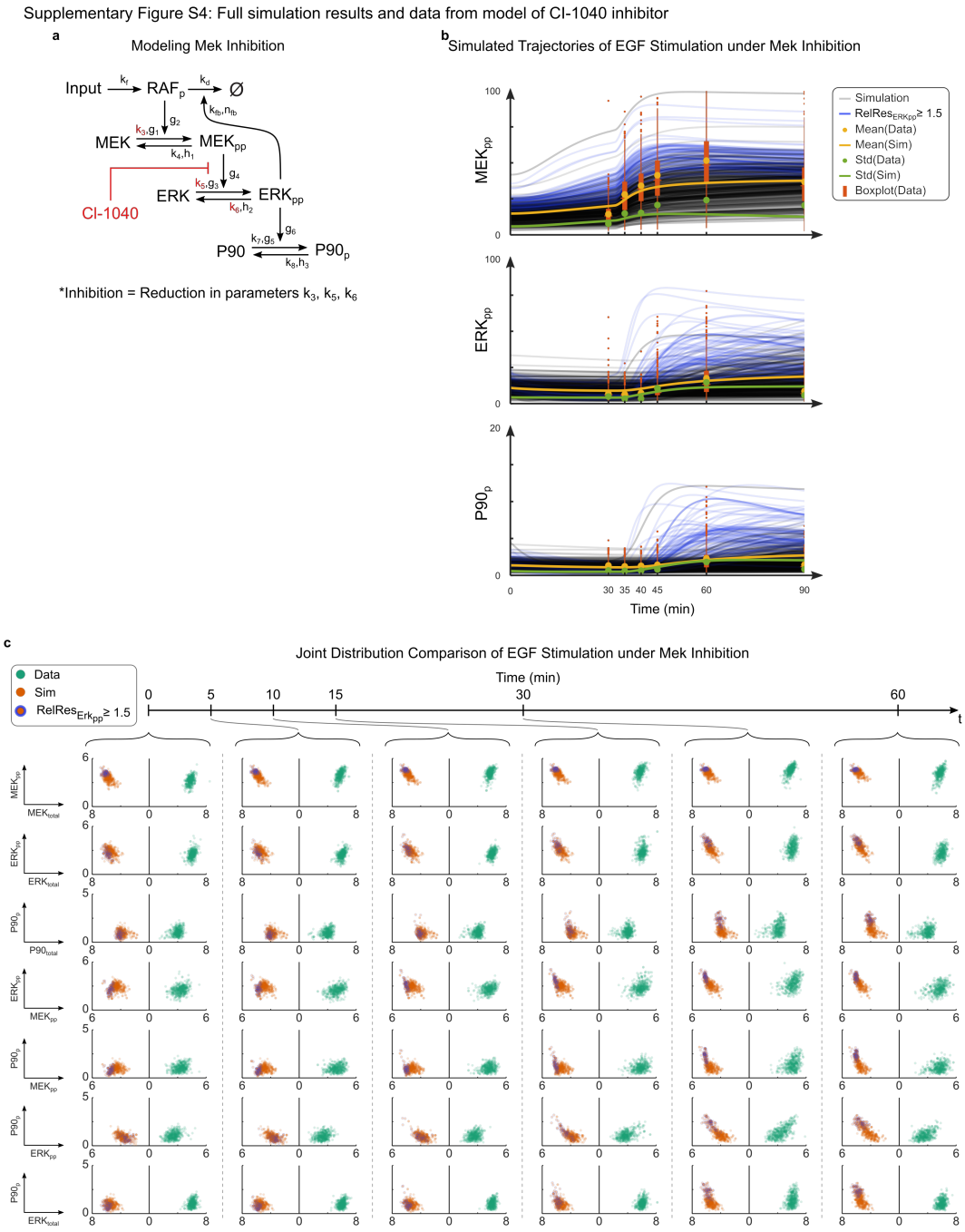


Figure B.4: Full simulation results of CI-1040 inhibitor. (**a**) Full model description of CI-1040 inhibitor after fitting to experimental data. Inhibitor acts on parameters in red. (**b**) Simulated single-cell trajectories of cells stimulated with EGF after 30 minute pre-treatment with CI-1040. **(c)** Simulated and measured snapshots of the same. Model units: scaled concentrations.

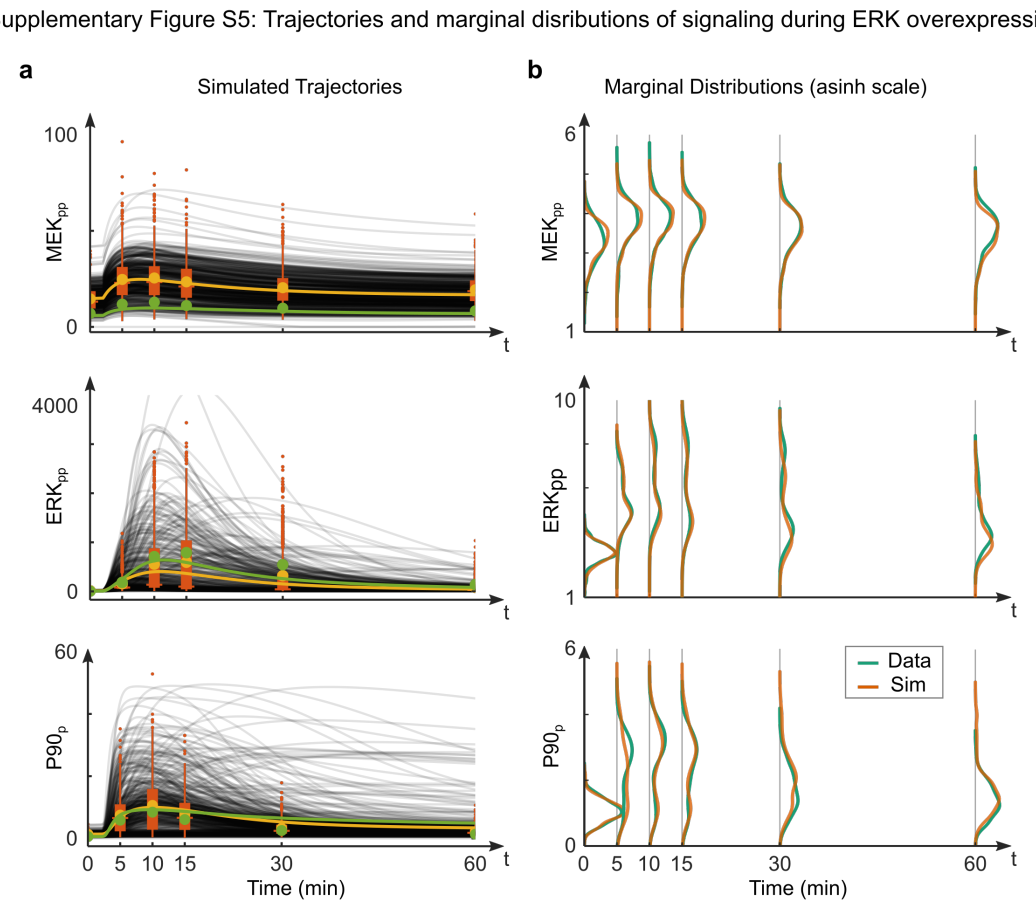


Figure B.5: Additional views of ERK overexpression simulations from Figure 3.6 of the main text. (**a**) Simulated single-cell trajectories for active proteins compared to measured population statistics. Yellow and green circles (or lines) are mean and standard deviation of the data (or the simulations) at each presented time point. Orange boxplots are middle 50 (bar) and 90 (line) percent of data with a horizontal line at median, while events outside 90% range are dots. (**b**) Marginal (one-dimensional) distribution of data (green) and simulation (orange). Model units: scaled concentrations.

Table B.4: Antibody panel. *Idu is not an antibody.

Channel	Element	Antibody	Clone
170	Er	RAS	Ras10
150	Nd	MEK	D1A5
144	Nd	pMEK1/2	166F8
143	Nd	ERK1/2	137F5
154	Sm	pERK1/2	20A
149	Sm	P90RSK2	D21B2
163	Dy	pP90RSK	D5D8
175	Lu	S6	54D2
171	Yb	pS6	N7-548
164	Dy	AKT	C67E7
153	Eu	pAKT	D9E
89	Yb	pHH3	HTA28
156	Gd	CYCLIN B1	GNS-11
127	I	Idu*	
172	Yb	cleaved PARP	F21-852
159	Tb	GAPDH	6C5

Table B.5: Average expression of total protein $\bar{x}_j$ in HEK cells (Boss et al., 2013) used for calculation of z_j.

Protein	Log_{10} Expression
ERK1/2	8.213269198
MEK1/2	8.344192782
RPS6	8.59
AKT1/2/3	7.325021898
RPS6K1/3	7.512659245
GAPDH	8.673333333
ERBB1	5.796666667

APPENDIX C

SUPPLEMENTARY MATERIALS: MECHANISTIC MODEL RECONCILES SIGNALING DYNAMICS ACROSS AN EPITHELIAL MESENCHYMAL TRANSITION

C.0.1 Data exclusion criteria

Antibody labeling and cell staining was optimized to minimize the number of cell events with zero or low ion counts in measurement channels used for modeling (e.g., total Erk). Furthermore, measured cell events with fewer than 5 ion counts in total protein channels used in the model were excluded. The choice of five counts was made as a trade-off between the number of discarded events, which increases as the threshold increases, and the uncertainty of single-cell measurements, which is a decreasing function of ion counts. This cutoff excluded fewer than 1% of cells per sample.

C.0.2 Data scaling for use in modeling

Mass cytometry provides relative values of protein abundance. Absolute values depend on factors such as antibody labeling efficiency, antibody staining and detection sensitivity. Although relative differences between measured protein abundances can often be subsumed into reaction parameters in a model, we also rescaled measured protein levels for use in modeling. As absolute values of protein abundance were not available for Py2T cells, total protein abundance values were taken to be 500 in arbitrary units. These values were used to estimate the average concentration $\bar{\mathbf{x}}_j$ of each protein j in a cell. Protein measurements were then linearly scaled as follows:

Given a protein j, its associated average concentration $\bar{\mathbf{x}}_j$ and its measured steady state distribution $D_{\mathbf{x}_{j|ss}}$, a scaling factor z_j was calculated as

$$z_j = \frac{\bar{\mathbf{x}}_j}{\mathrm{E}\big[D_{\mathbf{x}_{j|ss}}\big]}. \tag{11}$$

This implementation represents a linear scaling factor of the experimentally measured steady-state distribution of protein j such that the mean of the steady-state distribution $\mathrm{E}[D_{\mathbf{x}_{j|ss}}]$ is equal to the population average $\bar{\mathbf{x}}_j$ arbitrarily set at 500. For all experimental measurements, e.g., after a perturbation, each total protein j was scaled using the corresponding z_j.

Steady-state levels of phosphoproteins clearly cannot not be greater than those of the corresponding total protein pools. Additionally, if steady-state phosphoprotein levels are too high relative to total protein, the relative increase in phosphorylation levels in the model would be capped due to a lack of unphosphorylated protein. To avoid these issues, phosphoprotein distributions were scaled such that the average steady-state value of a phosphoprotein was a fractional value of the total protein level. The additional scaling factors for phophoproteins were set to 0.06. These values were chosen to reduce the number of cells in violation of the "active cannot be greater than total" constraint at steady state. In the special case that individual cells violated this constraint, these cells were excluded from the sample for analysis.

C.1 Model of EGF signaling in the MAPK/ERK cascade

In this section, we describe the mathematical formulation of our model of EGF signaling in the MAPK/ERK cascade.

C.1.1 State variables

The model uses eight state variables to describe changes in ERK pathway signaling components.

Table C.1: Model variables.

State variable	Description
I_1	Input to Raf
pRaf	Active Raf
Mek	Inactive Mek
ppMek	Active Mek
Erk	Inactive Erk
ppErk	Active Erk
Rsk	Inactive Rsk
pRsk	Active Rsk
I_2	Input to PI3K
PI3K	Active PI3K
Akt	Inactive Akt
pAkt	Active Akt
pGsk3b	Inactive Gsk3β
Gsk3b	Active Gsk3β
S6	Inactive S6
pS6	Active S6

Active and total Mek, Erk, Rsk, S6, Akt and Gsk3β were measured. After scaling, inactive forms were calculated as Inactive = Total − Active. The inputs I_1, I_2 and initial values of pRaf and PI3K were assumed to be the same for all cells. These choices were made to reduce prior assumptions. The variation in these components that were not explicitly measured were captured by the single-cell parameters k_d1, k_4, k_d2 and k_10, which were obtained from steady-state measurements (as in section B.2.2).

C.1.2 Kinetic parameters

Table C.2: Model parameters.

Parameter	Description
k_{f1}	Activation rate constant of pRaf by $Input_1$
g_{rac}	Kinetic order of crosstalk between PI3K and pRaf
k_{d1}	Degradation rate constant of pRaf signal
g_{D1}	Kinetic order of pRaf inactivation
k_3	Activation rate constant of ppMek
k_4	Inactivation rate constant of ppMek
g_1	Kinetic order of ppMek activation by Mek
g_2	Kinetic order of ppMek activation by pRAF
h_1	Kinetic order of ppMek inactivation by ppMek
k_5	Activation rate constant of ppErk
k_6	Inactivation rate constant of ppErk
g_3	Kinetic order of ppErk activation by Erk
g_4	Kinetic order of ppErk activation from ppMek
h_2	Kinetic order of ppErk inactivation from ppErk
k_7	Activation rate constant of pRsk
k_8	Inactivation rate constant of pRsk
g_5	Kinetic order of pP90 activation by Rsk
g_6	Kinetic order of pP90 activation by ppErk
h_3	Kinetic order of pP90 inactivation by pRsk
k_{f2}	Activation rate constant of PI3K by $Input_2$
k_{d2}	Degradation rate constant of PI3K signal
g_{D2}	Kinetic order of PI3K inactivation
k_{fb1}	Hill function inflection point of negative feedback from ppErk
k_9	Activation rate constant of pAkt
k_{10}	Inactivation rate constant of pAkt
g_7	Kinetic order of pAkt activation by Akt
g_8	Kinetic order of pAkt activation by PI3K
h_4	Kinetic order of pAkt inactivation by pAkt
k_{11}	Inctivation rate constant of Gsk3b
k_{12}	Activation rate constant of Gsk3b
g_9	Kinetic order of pGsk3b inactivation by Gsk3b
g_{10}	Kinetic order of pGsk3b inactivation by pRsk crosstalk
g_{11}	Kinetic order of pGsk3b inactivation by pAkt
h_5	Kinetic order of pGsk3b activation by pGsk3b
k_{13}	Activation rate constant of pS6
k_{14}	Inactivation rate constant of pS6
g_{12}	Kinetic order of pS6 activation by S6
g_{13}	Kinetic order of pS6 activation by pRsk
g_{14}	Kinetic order of pS6 activation by pAkt
h_6	Kinetic order of pS6 inactivation by pS6

The rate constants $\{k_d1, k_4.k_6, k_8, k_d2, k_10, k_12, k_14\}$ were in $\mathbf{\Phi_k}$ and computed from the steady-state equations as shown in section B.2.2. All other parameters were in $\mathbf{\Theta}$ and, therefore, equal across all cells and used as decision variables in the optimization algorithms.

C.1.3 Inputs

Table C.3: Model inputs.

Input	Description
I_1	Input (EGF) to pRaf
I_2	Input (EGF) to PI3K
I_{ss}	Input at steady state (before EGF addition: $t < \tau$)
τ_1	Time delay between addition of EGF and initial pRaf signaling
τ_2	Time delay between addition of EGF and initial PI3K signaling

The ERK/AKT pathway components considered in our model were in a pseudo-steady state at time scale of our experiments (one hour). Thus, our model must also be at steady state before simulated addition of EGF. We arbitrarily chose the input value of $I_1 = I_2 = I_{ss} = 1$ as the pre-stimulation steady state inputs to our model. We used the time delays τ_1 and τ_2 to represent the delay between experimental addition of EGF to the medium and the time when the signal reached the ERK and AKT signaling branches, respectively. In other words, τ_1 represents the time it takes for "signal" to pass be transmitted by reactions, such as receptor-ligand binding, receptor activation, etc., and reach Raf activation. The same for τ_2 and PI3K activation. Although the exact values of τ_1 and τ_2 for each cell are undoubtedly variable across the population, as not all cells will encounter the EGF signal at the same moment, we assumed each cell to have the same delays (τ) to maintain the entirely deterministic nature of our model. Thus, in the model for each $i \in \{1, 2\}$, if time $t < \tau_i$, then Input $I = I_{ss} = 1$. Once time $t = \tau_i$, then input is reset to represent EGF addition and Input $I = I_i$. At all times otherwise, input I_i is a dependent variable and determined by solution of the ODE system.

C.1.4 Equations

$$\dot{I_1} = -k_{f1} * (I_1 - I_{ss})$$

$$p\dot{R}af = k_{f1} * I_1 * PI3K^{g_{rac}} - k_{d1} * pRaf^{g_{D1}} * \frac{(ppErk^{f_n}}{k_{fb1}^{f_n} + ppErk^{f_n}}$$

$$\dot{Mek} = -k_3 * Mek^{g_1} * pRaf^{g_2} + k_4 * ppMek^{h_1}$$

$$pp\dot{M}ek = k_3 * Mek^{g_1} * pRaf^{g_2} - k_4 * ppMek^{h_1}$$

$$\dot{Erk} = -k_5 * Erk^{g_3} * ppMek^{g_4} + k_6 * ppErk^{h_2}$$

$$pp\dot{E}rk = k_5 * Erk^{g_3} * ppMek^{g_4} - k_6 * ppErk^{h_2}$$

$$\dot{Rsk} = -k_7 * Rsk^{g_5} * ppErk^{g_6} + k_8 * pRsk^{h_3}$$

$$p\dot{R}sk = k_7 * Rsk^{g_5} * ppErk^{g_6} - k_8 * pRsk^{h_3}$$

$$\dot{I_2} = -k_{f2} * (I_2 - I_{ss})$$

$$P\dot{I}3K = k_{f2} * I_2 - kd2 * pPI3K^{g_{D2}} * ppErk^{g_{fb2}}$$

$$\dot{Akt} = -k_9 * Akt^{g_7} * PI3K^{g_8} + k_{10} * pAkt^{h_4}$$

$$p\dot{A}kt = k_9 * Akt^{g_7} * PI3K^{g_8} - k_{10} * pAkt^{h_4}$$

$$Gs\dot{k}3b = -k_{11} * Gsk3b_{g_9} * pRsk^{g_{10}} * pAkt^{g_{11}} + k_{12} * pGsk3b^{h_5}$$

$$pGs\dot{k}3b = k_{11} * Gsk3b_{g_9} * pRsk^{g_{10}} * pAkt^{g_{11}} - k_{12} * pGsk3b^{h_5}$$

$$\dot{S6} = -k_{13} * S6^{g_{12}} * pRsk^{g_{13}} * pAkt^{g_{14}} + k_{14} * pS6^{h_6}$$

$$p\dot{S}6 = k_{13} * S6^{g_{12}} * pRsk^{g_{13}} * pAkt^{g_{14}} - k_{14} * pS6^{h_6}$$

C.2 Supplementary Figures and Tables

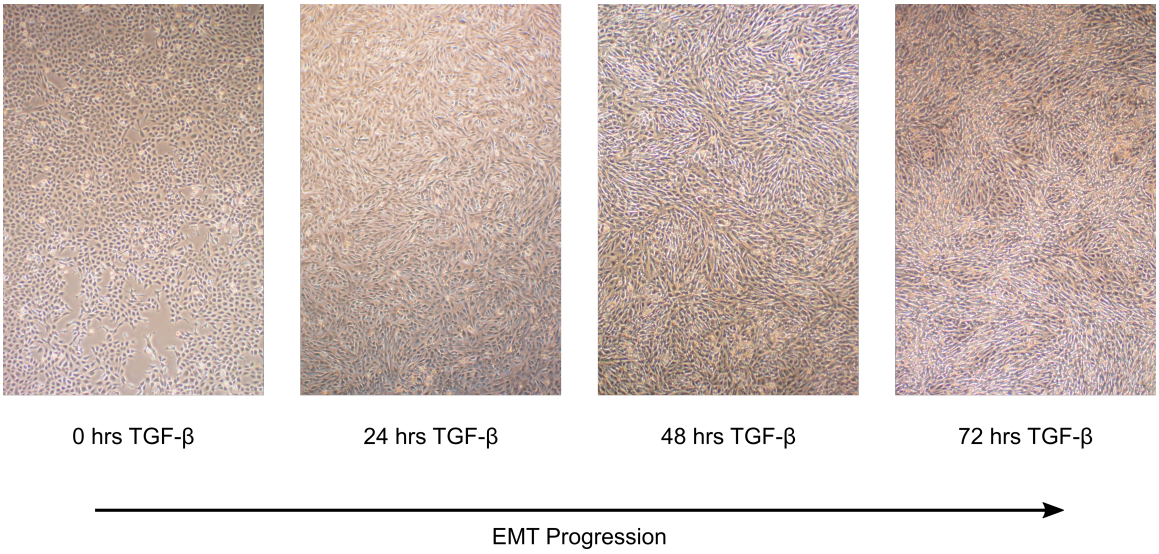


Figure C.1: Microscopy images of TGF-β treatment inducing an EMT. At day 0 all cells grow in a monolayer with a cobblestone epithelial morphology. By day 3, a subset of cells are no longer constrained to monolayer growth and have transitioned to an elongated mesenchymal morphology.

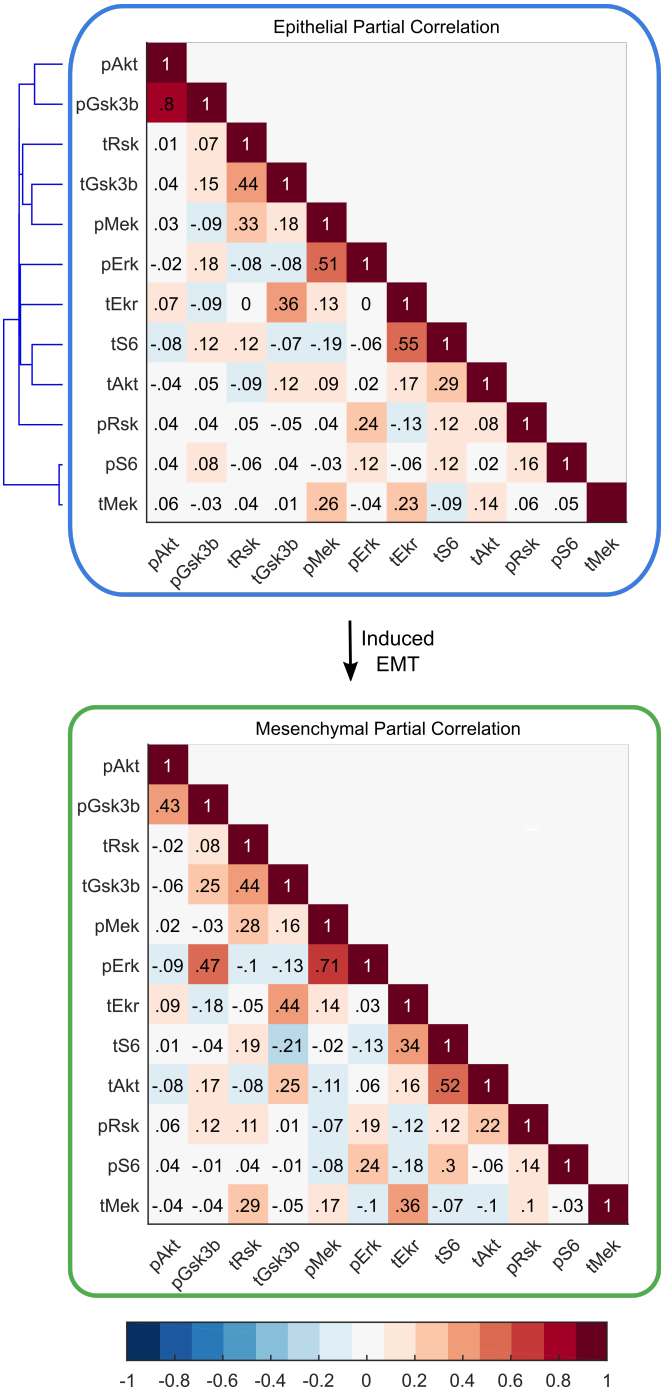


Figure C.2: Partial Correlation of phospho- and total signaling kinases. Partial correlation has been Fisher z-transformed. Both heatmaps are ordered by clustering the partial correlations of epithelial cells as shown in the dendrogram.

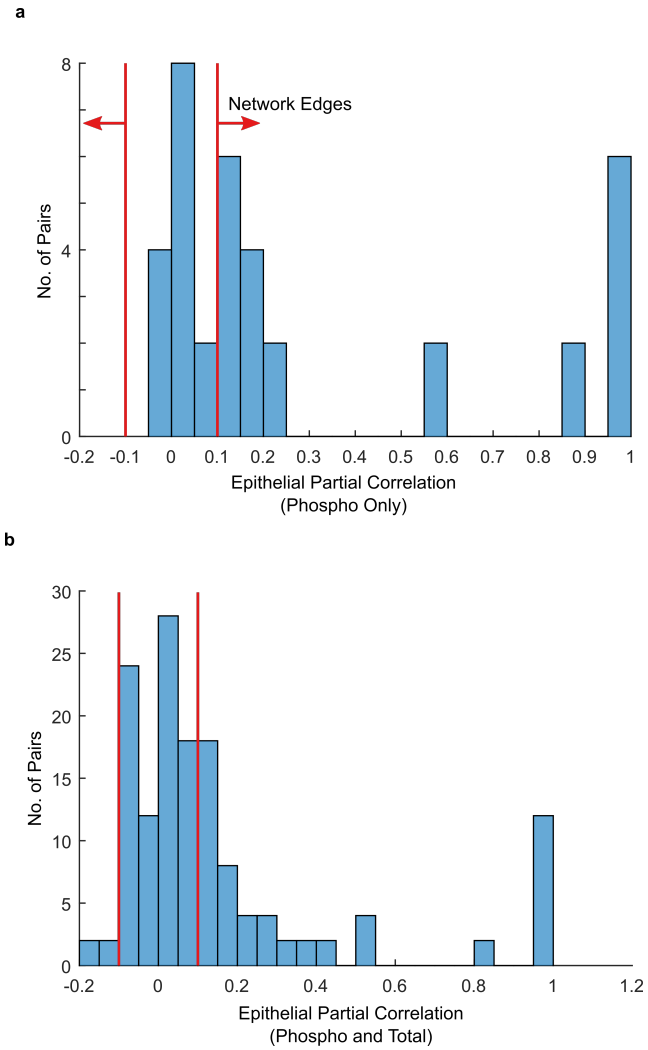


Figure C.3: Distribution of partial correlations. (**a**) All partial correlation values for epithelial cells as in Figure 4.2b. The cutoffs for network edges of |0.1| are marked by the red lines. (**b**) The distribution of partial correlation values for epithelial cells in Figure C.2. Red lines at |0.1| as in (a).

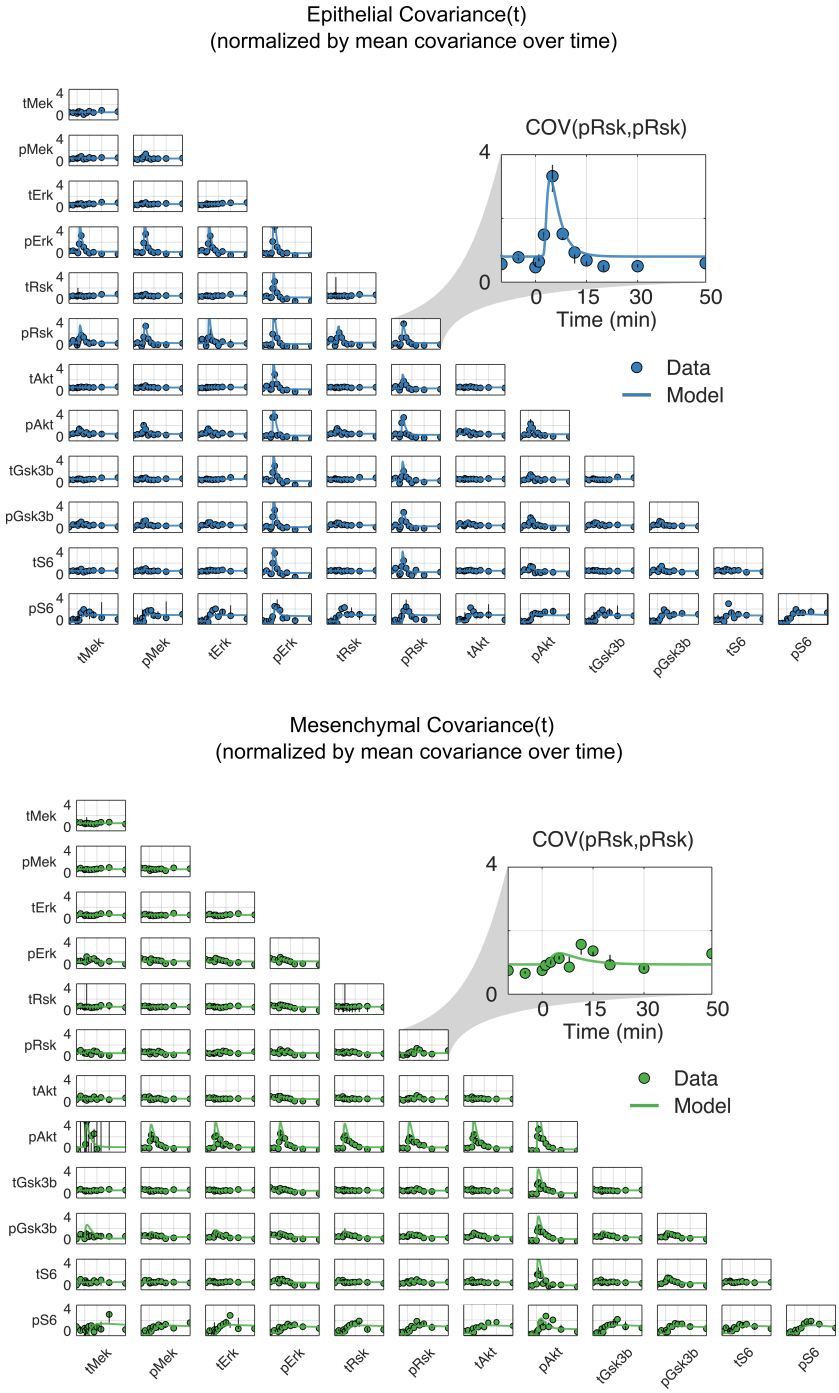


Figure C.4: Model and data covariance over time for all total and phospho pairs. Extended figure from Figure 4.3 to include total proteins. Dots are data. Colored lines are model. Black lines are range of data across replicates.

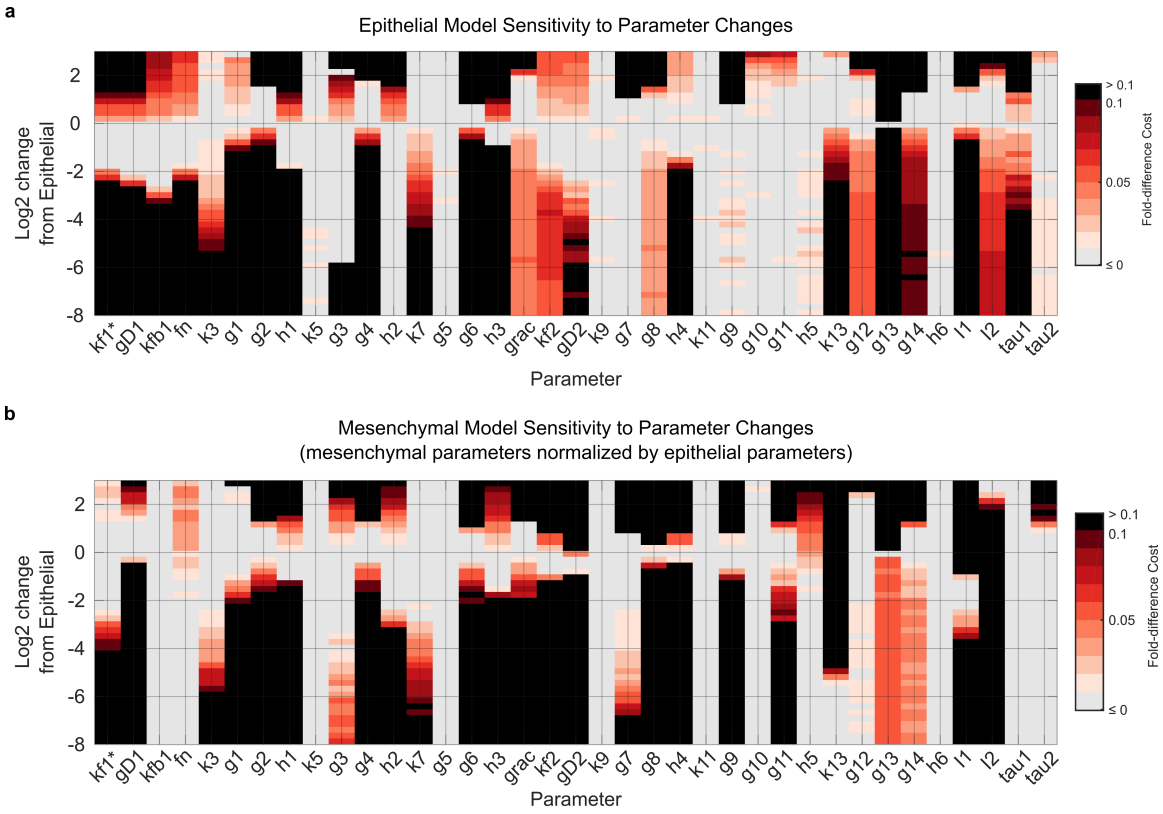


Figure C.5: Grid-based sensitivity of model to both cell phenotypes. Consideration of both phenotypes adds additional constraints to model parameters. (**a**) Model sensitivities calculated using epithelial cells as in Figure 4.4. (**b**) Model sensitivities calculated using Mesenchymal cells. Both plots show parameter values as $\log_2$ fold change from the optimal parameter values for epithelial cells.

Table C.4: Antibody Panel. Measured phosphoproteins that did not show robust dynamics in response to EGF stimulation were omitted from modeling.

Element	Antibody	Clone	[μg/mL]
139La	pCREB	J151-21	1.50
141Pr	pSTAT1	4a	2.00
142Nd	PTEN	138G6	0.75
143Nd	ERK1/2	137F5	1.00
144Nd	pMEK1/2	166F8	1.00
146Nd	pSTAT5	47	2.00
147Sm	GSk3β	D5C5Z	1.25
148Nd	pS6K	1A5	1.25
149Sm	P90RSK2	D21B2	0.75
150Nd	MEK	D1A5	2.25
151Eu	pEGFR	Y38	1.50
152Sm	pAMPKα	40H9	1.50
153Eu	pAKT	D9E	1.00
154Sm	pERK1/2	20A	0.75
155Gd	STAT1	SM1	2.00
156Gd	CYCLIN B1	GNS-11	1.00
158Gd	pGSK3	D85E12	0.25
159Tb	GAPDH	6C5	0.50
160Gd	mTOR	7C10	1.00
161Dy	pPDPK1	pS241	0.05
162Dy	Vimenten	D21H3	0.75
163Dy	pP90RSK	D5D8	1.25
164Dy	AKT	C67E7	0.50
165Ho	non-p-β-CATENIN	D13A1	0.75
166Er	pSTAT3	4/pSTAT3	1.50
167Er	STAT3	124H6	0.75
168Er	pPLCγ2	K86-689.37	1.00
169Tm	EGFR	AY13	1.00
170Er	pHH3	HTA28	0.50
171Yb	pS6	N7-548	0.10
172Yb	cleaved PARP	F21-852	2.00
173Yb	pMTOR	D9C2	2.00
174Yb	E-CADHERIN	36/E	0.50
175Lu	S6	54D2	0.25
176Yb	p4EBP1	236B4	0.25

Bibliography

Abu-Remaileh, M., Bender, S., Raddatz, G., Ansari, I., Cohen, D., Gutekunst, J., Musch, T., Linhart, H., Breiling, A., Pikarsky, E., Bergman, Y., and Lyko, F. (2015). Chronic inflammation induces a novel epigenetic program that is conserved in intestinal adenomas and in colorectal cancer. *Cancer Research*, 75(10):2120–2130.

Aldridge, B. B., Burke, J. M., Lauffenburger, D. a., and Sorger, P. K. (2006). Physicochemical modelling of cell signalling pathways. *Nature cell biology*, 8(11):1195–203.

Allen, L., Sebolt-Leopold, J., and Meyer, M. B. (2003). CI-1040 (PD184352), a targeted signal transduction inhibitor of MEK (MAPKK). *Seminars in Oncology*, 30(5 Suppl 16):105–116.

Altschuler, S. J. and Wu, L. F. (2010). Cellular Heterogeneity: Do Differences Make a Difference? *Cell*, 141(4):559–563.

Anjum, R. and Blenis, J. (2008). The RSK family of kinases: emerging roles in cellular signalling. *Nature Reviews Molecular Cell Biology*, 9(10):747–758.

Aoki, K. (2013). Stochastic erk activation induced by noise and cell-to-cell propagation regulates cell density-dependent proliferation. *Mol. Cell*, 52:529–540.

Bandura, D. R., Baranov, V. I., Ornatsky, O. I., Antonov, A., Kinach, R., Lou, X., Pavlov, S., Vorobiev, S., Dick, J. E., and Tanner, S. D. (2009). Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical chemistry*, 81(16):6813–22.

Bendall, S. C. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332:687–696.

Bendall, S. C., Nolan, G. P., Roederer, M., and Chattopadhyay, P. K. (2012). A deep profiler's guide to cytometry. *Trends in immunology*, 33(7):323–32.

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E. A. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., Balderas, R. S., Plevritis, S. K., Sachs, K., Pe'er, D., Tanner, S. D., and Nolan, G. P. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696.

Blume-Jensen, P. and Hunter, T. (2001). Oncogenic kinase signalling. *Nature*, 411(6835):355–65.

Bodenmiller, B., Zunder, E. R., Finck, R., Chen, T. J., Savig, E. S., Bruggner, R. V., Simonds, E. F., Bendall, S. C., Sachs, K., Krutzik, P. O., and Nolan, G. P. (2012). Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nature Biotechnology*, 30(9):857–866.

Boss, D., Kühn, J., Jourdain, P., Depeursinge, C., Magistretti, P. J., and Marquet, P. (2013). Measurement of absolute cell volume, osmotic membrane water permeability, and refractive index of transmembrane water and solute flux by digital holographic microscopy. *Journal of Biomedical Optics*, 18(3):036007.

Bowman, T., Garcia, R., Turkson, J., and Jove, R. (2000). Stats in oncogenesis. *Oncogene*, 19:2474–2488.

Brightman, F. A. and Fell, D. A. (2000). Differential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cells. *FEBS Letters*, 482(3):169–174.

Bronstein, L., Zechner, C., and Koeppl, H. (2015). Bayesian inference of reaction kinetics from single-cell recordings across a heterogeneous cell population. *Methods*, 85:22–35.

Bunt, G. and Wouters, F. S. (2017). FRET from single to multiplexed signaling events. *Biophysical Reviews*, 9(2):119–129.

Burrell, R. A. and Swanton, C. (2014). Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular Oncology*, 8(6):1095–1111.

Byers, L. A., Diao, L., Wang, J., Saintigny, P., Girard, L., Peyton, M., Shen, L., Fan, Y., Giri, U., Tumula, P. K., Nilsson, M. B., Gudikote, J., Tran, H., Cardnell, R. J., Bearss, D. J., Warner, S. L., Foulks, J. M., Kanner, S. B., Gandhi, V., Krett, N., Rosen, S. T., Kim, E. S., Herbst, R. S., Blumenschein, G. R., Lee, J. J., Lippman, S. M., Ang, K. K., Mills, G. B., Hong, W. K., Weinstein, J. N., Wistuba, I. I., Coombes, K. R., Minna, J. D., and Heymach, J. V. (2013). An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clinical Cancer Research*, 19(1):279–290.

Bywater, M. J., Pearson, R. B., McArthur, G. A., and Hannan, R. D. (2013). Dysregulation of the basal rna polymerase transcription apparatus in cancer. *Nat. Rev. Cancer*, 13:299–314.

Cardaci, S., Filomeni, G., and Ciriolo, M. R. (2012). Redox implications of ampk-mediated signal transduction beyond energetic clues. *J. Cell Sci.*, 125:2115–2125.

Caunt, C. J., Sale, M. J., Smith, P. D., and Cook, S. J. (2015). MEK1 and MEK2 inhibitors and cancer therapy: The long and winding road. *Nature Reviews Cancer*, 15(10):577–592.

Cepero, V. (2010). Met and kras gene amplification mediates acquired resistance to met tyrosine kinase inhibitors. *Cancer Res.*, 70:7580–7590.

Chang-Yew Leow, C., Gerondakis, S., and Spencer, A. (2013). MEK inhibitors as a chemotherapeutic intervention in multiple myeloma. *Blood Cancer Journal*, 3(3):e105.

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090–aaa6090.

Chevrier, S., Levine, J. H., Zanotelli, V. R. T., Silina, K., Schulz, D., Bacac, M., Ries, C. H., Ailles, L., Jewett, M. A. S., Moch, H., van den Broek, M., Beisel, C., Stadler, M. B., Gedye, C., Reis, B., Pe'er, D., and Bodenmiller, B. (2017). An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell*, 169(4):736–749.e18.

Chung, K., Rivet, C. A., Kemp, M. L., and Lu, H. (2011). Imaging single-cell signaling dynamics with a deterministic high-density single-cell trap array. *Analytical Chemistry*, 83(18):7044–7052.

Citri, A. and Yarden, Y. (2006). EGF-ERBB signalling: Towards the systems level.

Cohen, a. a., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M., Issaeva, I., Sigal, a., Milo, R., Cohen-Saidon, C., Liron, Y., Kam, Z., Cohen, L., Danon, T., Perzov, N., and Alon, U. (2008). Dynamic proteomics of individual cancer cells in response to a drug. *Science (New York, N.Y.)*, 322(5907):1511–6.

Cohen, P. and Frame, S. (2001). The renaissance of gsk3. *Nat. Rev. Mol. Cell Biol.*, 2:769–776.

Cohen-Saidon, C., Cohen, A. A., Sigal, A., Liron, Y., and Alon, U. (2009). Dynamics and Variability of ERK2 Response to EGF in Individual Living Cells. *Molecular Cell*, 36(5):885–893.

Corcoran, R. B. (2013). Synthetic lethal interaction of combined bcl-xl and mek inhibition promotes tumor regressions in kras mutant cancer models. *Cancer Cell*, 23:121–128.

Couzens, A. L., Knight, J. D. R., Kean, M. J., Teo, G., Weiss, A., Dunham, W. H., Lin, Z.-Y., Bagshaw, R. D., Sicheri, F., Pawson, T., Wrana, J. L., Choi, H., and Gingras, A.-C. (2013). Protein Interaction Network of the Mammalian Hippo Pathway Reveals Mechanisms of Kinase-Phosphatase Interactions. *Science Signaling*, 6(302):rs15–rs15.

Creixell, P., Schoof, E. M., Simpson, C. D., Longden, J., Miller, C. J., Lou, H. J., Perryman, L., Cox, T. R., Zivanovic, N., Palmeri, A., Wesolowska-Andersen, A., Helmer-Citterich, M., Ferkinghoff-Borg, J., Itamochi, H., Bodenmiller, B., Erler, J. T., Turk, B. E., and Linding, R. (2015). Kinome-wide Decoding of Network-Attacking Mutations Rewiring Cancer Signaling. *Cell*, 163(1):202–217.

Dagogo-Jack, I. and Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies.

Davies, H. (2002). Mutations of the braf gene in human cancer. *Nature*, 417:949–954.

De Los Angeles, A. (2015). Hallmarks of pluripotency. *Nature*, 525:469–478.

Desai, P., Yang, J., Tian, B., Sun, H., Kalita, M., Ju, H., Paulucci-Holthauzen, A., Zhao, Y., Brasier, A. R., and Sadygov, R. G. (2015). Mixed-effects model of epithelial-mesenchymal transition reveals rewiring of signaling networks. *Cellular Signalling*, 27(7):1413–1425.

Dhar, S. S., Chakraborty, B., and Chaudhuri, P. (2014). Comparison of multivariate distributions using quantilequantile plots and related tests. *Bernoulli*, 20(3):1484–1506.

Dolmetsch, R. E., Lewis, R. S., Goodnow, C. C., and Healy, J. I. (1997). Differential activation of transcription factors induced by Ca2+ response amplitude and duration. *Nature*, 386(6627):855–858.

Du, B. and Shim, J. S. (2016). Targeting epithelial-mesenchymal transition (EMT) to overcome drug resistance in cancer.

Ebi, H., Costa, C., Faber, A. C., Nishtala, M., Kotani, H., Juric, D., Della Pelle, P., Song, Y., Yano, S., Mino-Kenudson, M., Benes, C. H., and Engelman, J. A. (2013). PI3K regulates MEK/ERK signaling in breast cancer via the Rac-GEF, P-Rex1. *Proceedings of the National Academy of Sciences*, 110(52):21124–21129.

Egea, J. A., Henriques, D., Cokelaer, T., Villaverde, A. F., MacNamara, A., Danciu, D.-P., Banga, J. R., and Saez-Rodriguez, J. (2014). MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinformatics*, 15(1):136.

Eralp, Y., Derin, D., Ozluk, Y., Yavuz, E., Guney, N., Saip, P., Muslumanoglu, M., Igci, A., Kücücük, S., Dincer, M., Aydiner, A., and Topuz, E. (2008a). MAPK overexpression is associated with anthracycline resistance and increased risk for recurrence in patients with triple-negative breast cancer. *Annals of Oncology*, 19(4):669–674.

Eralp, Y., Derin, D., Ozluk, Y., Yavuz, E., Guney, N., Saip, P., Muslumanoglu, M., Igci, A., Kücücük, S., Dincer, M., Aydiner, A., and Topuz, E. (2008b). MAPK overexpression is associated with anthracycline resistance and increased risk for recurrence in patients with triple-negative breast cancer. *Annals of Oncology*, 19(4):669–674.

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Reports*, 10(8):1386–1397.

Feinberg, A. P. (2007). Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447:433–440.

Ferrell, J. E. (2016). Perfect and near-perfect adaptation in cell signaling.

Filippi, S., Barnes, C. P., Kirk, P. D. W., Kudo, T., Kunida, K., McMahon, S. S., Tsuchiya, T., Wada, T., Kuroda, S., and Stumpf, M. P. H. (2016). Robustness of MEK-ERK Dynamics and Origins of Cell-to-Cell Variability in MAPK Signaling. *Cell Reports*, 15(11):2524–2535.

Finck, R., Simonds, E. F., Jager, A., Krishnaswamy, S., Sachs, K., Fantl, W., Pe'er, D., Nolan, G. P., and Bendall, S. C. (2013). Normalization of mass cytometry data with bead standards. *Cytometry Part A*, 83 A(5):483–494.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annals of Statistics*, 7(4):697–717.

Fu, D.-J., Miller, A. D., Southard, T. L., Flesken-Nikitin, A., Ellenson, L. H., and Yu Nikitin, A. (2018). Stem Cell Pathology. *Annual Review of Pathology: Mechanisms of Disease*, 13:71–92.

Garmaroudi, F. S., Marchant, D., Si, X., Khalili, A., Bashashati, A., Wong, B. W., Tabet, A., Ng, R. T., Murphy, K., Luo, H., Janes, K. A., and McManus, B. M. (2010). Pairwise network mechanisms in the host signaling response to coxsackievirus B3 infection. *Proceedings of the National Academy of Sciences*, 107(39):17053–17058.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Molecular & Cellular Proteomics*, 11(3):M111.014050.

Giesen, C., Wang, H. a. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., Varga, Z., Wild, P. J., Günther, D., and Bodenmiller, B. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417–22.

Gillies, T. E., Pargett, M., Minguet, M., Davies, A. E., and Albeck, J. G. (2017). Linear Integration of ERK Activity Predominates over Persistence Detection in Fra-1 Regulation. *Cell Systems*, 5(6):549–563.e5.

Goutsias, J. (2007). Classical versus stochastic kinetics modeling of biochemical reaction systems. *Biophysical Journal*, 92(7):2350–2365.

Govindarajan, B., Sligh, J. E., Vincent, B. J., Li, M., Canter, J. A., Nickoloff, B. J., Rodenburg, R. J., Smeitink, J. A., Oberley, L., Zhang, Y., Slingerland, J., Arnold, R. S., Lambeth, J. D., Cohen, C., Hilenski, L., Griendling, K., Martínez-Diez, M., Cuezva, J. M., and Arbiser, J. L. (2007). Overexpression of Akt converts radial growth melanoma to vertical growth melanoma. *Journal of Clinical Investigation*, 117(3):719–729.

Grant, D. M., Zhang, W., McGhee, E. J., Bunney, T. D., Talbot, C. B., Kumar, S., Munro, I., Dunsby, C., Neil, M. a. a., Katan, M., and French, P. M. W. (2008). Multiplexed FRET to image multiple signaling events in live cells. *Biophysical journal*, 95(10):L69–71.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A Kernel Two-sample Test. *J. Mach. Learn. Res.*, 13:723–773.

Gretton, A., Sriperumbudur, B., Sejdinovic, D., Strathmann, H., and Pontil, M. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Neural Information Processing Systems*, pages 1214–1222.

Gut, G., Herrmann, M. D., and Pelkmans, L. (2018). Multiplexed protein maps link subcellular organization to cellular states. *Science*, 361(6401).

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. *Proc. 7th Python Sci. Conf.*, 2008:11–15.

Halasz, M., Kholodenko, B. N., Kolch, W., and Santra, T. (2016). Integrating network reconstruction with mechanistic modeling to predict cancer therapies. *Science Signaling*, 9(455):ra114.

Han, T., Xiang, D. M., Sun, W., Liu, N., Sun, H. L., Wen, W., Shen, W. F., Wang, R. Y., Chen, C., Wang, X., Cheng, Z., Li, H. Y., Wu, M. C., Cong, W. M., Feng, G. S., Ding, J., and Wang, H. Y. (2015). PTPN11/Shp2 overexpression enhances liver cancer progression and predicts poor prognosis of patients. *Journal of Hepatology*, 63(3):651–660.

Hasenauer, J., Hasenauer, C., Hucho, T., and Theis, F. J. (2014). ODE Constrained Mixture Modelling: A Method for Unraveling Subpopulation Structures and Dynamics. *PLoS Computational Biology*, 10(7).

Hasenauer, J., Waldherr, S., Doszczak, M., Radde, N., Scheurich, P., and Allgöwer, F. (2011). Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, 12(1):125.

Hendriks, R. W., Yuvaraj, S., and Kil, L. P. (2014). Targeting bruton's tyrosine kinase in b cell malignancies. *Nat. Rev. Cancer*, 14:219–232.

Hengl, S., Kreutz, C., Timmer, J., and Maiwald, T. (2007). Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612–2618.

Hill, S. M., Nesser, N. K., Johnson-Camacho, K., Jeffress, M., Johnson, A., Boniface, C., Spencer, S. E., Lu, Y., Heiser, L. M., Lawrence, Y., Pande, N. T., Korkola, J. E., Gray, J. W., Mills, G. B., Mukherjee, S., and Spellman, P. T. (2017). Context Specificity in Causal Signaling Networks Revealed by Phosphoprotein Profiling. *Cell Systems*, 4(1):73–83.e10.

Huang, C. Y. and Ferrell, J. E. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19):10078–83.

Hughes, A. J., Spelke, D. P., Xu, Z., Kang, C.-C., Schaffer, D. V., and Herr, A. E. (2014). Single-cell western blotting. *Nature methods*, 11(7):749–55.

Jeong, H. M., Kwon, M. J., and Shin, Y. K. (2014). Overexpression of Cancer-Associated Genes via Epigenetic Derepression Mechanisms in Gynecologic Cancer. *Frontiers in Oncology*, 4:12.

Justel, A., Peña, D., and Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, 35(3):251–259.

Kholodenko, B. N. (2006). Cell-signalling dynamics in time and space. *Nature Reviews Molecular Cell Biology*, 7(3):165–76.

Kim, D., Rath, O., Kolch, W., and Cho, K.-H. (2007). A hidden oncogenic positive feedback loop caused by crosstalk between wnt and erk pathways. *Oncogene*, 26:4571–4579.

Kim, S. Y. (2012). Amp-activated protein kinase-a1 as an activating kinase of tgf-b-activated kinase 1 has a key role in inflammatory signals. *Cell Death Dis.*, 3:e357.

Kolitz, S. E. and Lauffenburger, D. a. (2012). Measurement and modeling of signaling at the single-cell level. *Biochemistry*, 51(38):7433–43.

Komatsu, N., Aoki, K., Yamada, M., Yukinaga, H., Fujita, Y., Kamioka, Y., and Matsuda, M. (2011). Development of an optimized backbone of FRET biosensors for kinases and GTPases. *Molecular Biology of the Cell*, 22(23):4647–4656.

Kramer, B. W., Götz, R., and Rapp, U. R. (2004). Use of mitogenic cascade blockers for treatment of C-Raf induced lung adenoma in vivo: CI-1040 strongly reduces growth and improves lung structure. *BMC Cancer*, 4(1):24.

Krishnaswamy, S. (2014). Systems biology. conditional density-based analysis of t cell signaling in single-cell data. *Science*, 346:1250689.

Krishnaswamy, S., Zivanovic, N., Sharma, R., Pe'er, D., and Bodenmiller, B. (2018). Learning time-varying information flow from single-cell epithelial to mesenchymal transition data. *PLOS ONE*, 13(10):e0203389.

Lee, M. J., Ye, A. S., Gardino, A. K., Heijink, A. M., Sorger, P. K., MacBeath, G., and Yaffe, M. B. (2012). Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell*, 149(4):780–794.

Lee, Y.-K., Hwang, J.-T., Kwon, D. Y., Surh, Y.-J., and Park, O. J. (2010). Induction of apoptosis by quercetin is mediated through ampkalpha1/ask1/p38 pathway. *Cancer Lett.*, 292:228–236.

Levchenko, A., Bruck, J., and Sternberg, P. W. (2000). Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. *Proceedings of the National Academy of Sciences*, 97(11):5818–5823.

Lin, J.-R., Fallahi-Sichani, M., and Sorger, P. K. (2015). Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nature Communications*, 6:8390.

Little, A. S. (2011). Amplification of the driving oncogene, kras or braf, underpins acquired resistance to mek1/2 inhibitors in colorectal cancer cells. *Sci. Signal.*, 4:ra17.

Loos, C., Moeller, K., Fröhlich, F., Hucho, T., and Hasenauer, J. (2018). A Hierarchical, Data-Driven Approach to Modeling Single-Cell Populations Predicts Latent Causes of Cell-To-Cell Variability. *Cell Systems*, 6(5):593–603.e13.

Lun, X.-K., Zanotelli, V. R. T., Wade, J. D., Schapiro, D., Tognetti, M., Dobberstein, N., and Bodenmiller, B. (2017). Influence of node abundance on signaling network state and dynamics analyzed by mass cytometry. *Nature Biotechnology*, 35(2):164–172.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Manning, B. D. and Cantley, L. C. (2007). Akt/pkb signaling: navigating downstream. *Cell*, 129:1261–1274.

Marin, T. M. (2008). Shp2 negatively regulates growth in cardiomyocytes by controlling focal adhesion kinase/src and mtor pathways. *Circ. Res.*, 103:813–824.

Massague, J. (2003). Integration of smad and mapk pathways: a link and a linker revisited. *Genes Dev.*, 17:2993–2997.

Meacham, C. E. and Morrison, S. J. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–37.

Mendoza, M. C., Er, E. E., and Blenis, J. (2011). The ras-erk and pi3k-mtor pathways: cross-talk and compensation. *Trends Biochem. Sci.*, 36:320–328.

Mitra, S. K., Hanson, D. A., and Schlaepfer, D. D. (2005). Focal adhesion kinase: in command and control of cell motility. *Nat. Rev. Mol. Cell Biol.*, 6:56–68.

Nyati, M. K., Morgan, M. A., Feng, F. Y., and Lawrence, T. S. (2006). Integration of egfr inhibitors with radiochemotherapy. *Nat. Rev. Cancer*, 6:876–885.

Olayioye, M. A., Neve, R. M., Lane, H. A., and Hynes, N. E. (2000). The erbb signaling network: receptor heterodimerization in development and cancer. *Embo J.*, 19:3159–3167.

Oliva, J. L., Griner, E. M., and Kazanietz, M. G. (2005). Pkc isozymes and diacylglycerol-regulated proteins as effectors of growth factor receptors. *Growth Factors*, 23:245–252.

Papin, J. a., Hunter, T., Palsson, B. O., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nature reviews. Molecular cell biology*, 6(2):99–111.

Perfetto, L. (2016). Signor: a database of causal relationships between biological entities. *Nucleic Acids Res.*, 44:D548–D554.

Perfetto, S. P., Chattopadhyay, P. K., and Roederer, M. (2004). Seventeen-colour flow cytometry: Unravelling the immune system.

Petsalaki, E., Helbig, A. O., Gopal, A., Pasculescu, A., Roth, F. P., and Pawson, T. (2015). SELPHI: Correlation-based identification of kinase-associated networks from global phospho-proteomics data sets. *Nucleic Acids Research*, 43(W1):W276–W282.

Rahman, M. T. (2013). Kras and mapk1 gene amplification in type ii ovarian carcinomas. *Int. J. Mol. Sci.*, 14:13748–13762.

Rapsomaniki, M. A., Lun, X.-K., Woerner, S., Laumanns, M., Bodenmiller, B., and Martínez, M. R. (2018). CellCycleTRACER accounts for cell cycle and volume in mass cytometry data. *Nature Communications*, 9(1):632.

Rawlings, J. S., Rosler, K. M., and Harrison, D. A. (2004). The jak/stat signaling pathway. *J. Cell Sci.*, 117:1281–1283.

Redell, M. S. (2013). Facs analysis of stat3/5 signaling reveals sensitivity to g-csf and il-6 as a significant prognostic factor in pediatric aml: a children's oncology group report. *Blood*, 121:1083–1093.

Regot, S., Hughey, J. J., Bajar, B. T., Carrasco, S., and Covert, M. W. (2014). High-sensitivity measurements of multiple kinase activities in live single cells. *Cell*, 157(7):1724–34.

Roberts, P. J. and Der, C. J. (2007). Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene*, 26(22):3291–3310.

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(4):515–530.

Roux, P. P., Richards, S. A., and Blenis, J. (2003). Phosphorylation of p90 Ribosomal S6 Kinase (RSK) Regulates Extracellular Signal-Regulated Kinase Docking and RSK Activity. *Molecular and Cellular Biology*, 23(14):4796–4804.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). Rubner, Tomasi, Guibas - 2000 - The Earth Mover ' s Distance as a Metric for Image Retrieval.pdf.

Ryu, H., Chung, M., Dobrzynski, M., Fey, D., Blum, Y., Lee, S. S., Peter, M., Kholodenko, B. N., Jeon, N. L., and Pertz, O. (2015). Frequency modulation of ERK activation dynamics rewires cell fate. *Molecular Systems Biology*, 11(11):838–838.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. a., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–9.

Salt, M. B., Bandyopadhyay, S., and McCormick, F. (2014). Epithelial-to-mesenchymal transition rewires the molecular path to PI3K-dependent proliferation. *Cancer Discovery*, 4(2):186–199.

Samatar, A. A. and Poulikakos, P. I. (2014). Targeting RASERK signalling in cancer: promises and challenges. *Nature Reviews Drug Discovery*, 13(12):928–942.

Santarius, T., Shipley, J., Brewer, D., Stratton, M. R., and Cooper, C. S. (2010). A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer*, 10:59–64.

Santos, S. D. M., Verveer, P. J., and Bastiaens, P. I. H. (2007). Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nature Cell Biology*, 9(3):324–330.

Savageau, M. A. (1976). *Biochemical systems analysis : a study of function and design in molecular biology*. Addison-Wesley Pub. Co., Advanced Book Program.

Sebolt-Leopold, J. S., Dudley, D. T., Herrera, R., Van Becelaere, K., Wiland, A., Gowan, R. C., Tecle, H., Barrett, S. D., Bridges, A., Przybranowski, S., Leopold, W. R., and Saltiel, A. R. (1999). Blockade of the MAP Kinase Pathway Suppresses Growth of Colon Tumors in Vivo. *Nature Medicine*, 5(7):810–816.

Selimkhanov, J., Taylor, B., Yao, J., Pilko, A., Albeck, J., Hoffmann, A., Tsimring, L., and Wollman, R. (2014). Accurate information transmission through dynamic biochemical signaling networks. *Science*, 346(6215):1370–1373.

Seong, H.-A., Jung, H., Ichijo, H., and Ha, H. (2010). Reciprocal negative regulation of pdk1 and ask1 signaling by direct interaction and phosphorylation. *J. Biol. Chem.*, 285:2397–2414.

Sever, R. and Brugge, J. S. (2015). Signal transduction in cancer. *Cold Spring Harbor Perspectives in Medicine*, 5(4).

Shaffer, S. M., Dunagin, M. C., Torborg, S. R., Torre, E. A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P. A., Xiao, M., Eggan, E., Anastopoulos, I. N., Vargas-Garcia, C. A., Singh, A., Nathanson, K. L., Herlyn, M., and Raj, A. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658):431–435.

Shaul, Y. D. and Seger, R. (2007). The MEK/ERK cascade: From signaling specificity to diverse functions. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1773(8):1213–1226.

Shibue, T. and Weinberg, R. A. (2017). EMT, CSCs, and drug resistance: The mechanistic link and clinical implications.

Shin, I., Kim, S., Song, H., Kim, H.-R. C., and Moon, A. (2005). H-ras-specific activation of rac-mkk3/6-p38 pathway: its critical role in invasion and migration of breast epithelial cells. *J. Biol. Chem.*, 280:14675–14683.

Silvera, D., Formenti, S. C., and Schneider, R. J. (2010). Translational control in cancer. *Nat. Rev. Cancer*, 10:254–266.

Snijder, B. and Pelkmans, L. (2011). Origins of regulated cell-to-cell variability. *Nature Reviews Molecular Cell Biology*, 12(2):119–125.

Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M., and Sorger, P. K. (2009). Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature*, 459(7245):428–32.

Spiller, D. G., Wood, C. D., Rand, D. A., and White, M. R. H. (2010). Measurement of single-cell dynamics. *Nature*, 465(7299):736–745.

Spitzer, M. H. and Nolan, G. P. (2016). Mass Cytometry: Single Cells, Many Features. *Cell*, 165(4):780–791.

Sundqvist, A. (2013). Specific interactions between smad proteins and ap-1 components determine tgfb-induced breast cancer cell invasion. *Oncogene*, 32:3606–3615.

Tebbutt, N., Pedersen, M. W., and Johns, T. G. (2013). Targeting the erbb family in cancer: couples therapy. *Nat. Rev. Cancer*, 13:663–673.

Tewari, M. (2004). Systematic interactome mapping and genetic perturbation analysis of a c. elegans tgf-b signaling network. *Mol. Cell*, 13:469–482.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386.

Ulianov, A., Müntener, O., and Schaltegger, U. (2015). The ICPMS signal as a Poisson process: a review of basic concepts. *J. Anal. At. Spectrom.*, 30(6):1297–1321.

Valtorta, E. (2013). Kras gene amplification in colorectal cancer and impact on response to egfr-targeted therapy. *Int. J. Cancer*, 133:1259–1265.

Van der Merwe, R. (2004). Sigma-point Kalman filters for probabilistic inference in dynamic state-space models. *PhD thesis*, (April):378.

Vanlier, J., Tiemann, C. A., Hilbers, P. A. J., and van Riel, N. A. W. (2012). An integrated strategy for prediction uncertainty analysis. *Bioinformatics*, 28(8):1130–1135.

Voit, E. O. (2000). *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists*. Cambridge University Press, UK.

Voit, E. O. (2013). Biochemical Systems Theory: A Review. *ISRN Biomathematics*, 2013:1–53.

Waldmeier, L., Meyer-Schaller, N., Diepenbruck, M., and Christofori, G. (2012). Py2T Murine Breast Cancer Cells, a Versatile Model of TGFβ-Induced EMT In Vitro and In Vivo. *PLoS ONE*, 7(11).

Wang, D. and Bodovitz, S. (2010). Single cell analysis: The new frontier in 'omics'.

Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, 15(18):3163–3168.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58:236–244.

Wee, P. and Wang, Z. (2017). Epidermal growth factor receptor cell proliferation signaling pathways.

Wei, Z. Z. (2014). Regulatory role of the jnk-stat1/3 signaling in neuronal differentiation of cultured mouse embryonic stem cells. *Cell. Mol. Neurobiol.*, 34:881–893.

Wellbrock, C. and Arozarena, I. (2016). The Complexity of the ERK/MAP-Kinase Pathway and the Treatment of Melanoma Skin Cancer. *Frontiers in Cell and Developmental Biology*, 4:33.

Will, T. and Helms, V. (2015). PPIXpress: Construction of condition-specific protein interaction networks based on transcript expression. *Bioinformatics*, 32(4):571–578.

Wolf-Yadlin, A. (2006). Effects of her2 overexpression on cell signaling networks governing proliferation and migration. *Mol. Syst. Biol.*, 2:54.

Xu, Y., Li, N., Xiang, R., and Sun, P. (2014). Emerging roles of the p38 mapk and pi3k/akt/mtor pathways in oncogene-induced senescence. *Trends Biochem. Sci.*, 39:268–276.

Yang, X., Boehm, J. S., Yang, X., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S. R., Alkan, O., Bhimdi, T., Green, T. M., Johannessen, C. M., Silver, S. J., Nguyen, C., Murray, R. R., Hieronymus, H., Balcha, D., Fan, C., Lin, C., Ghamsari, L., Vidal, M., Hahn, W. C., Hill, D. E., and Root, D. E. (2011). A public genome-scale lentiviral expression library of human ORFs. *Nature Methods*, 8(8):659–661.

Zhang, Y. (2005). Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Mol. Cell. Proteomics*, 4:1240–1250.

Zhou, Y. (2010). Chimeric mouse tumor models reveal differences in pathway activation between erbb family- and kras-dependent lung adenocarcinomas. *Nat. Biotechnol.*, 28:71–78.

Zunder, E. R., Finck, R., Behbehani, G. K., Amir, E. A. D., Krishnaswamy, S., Gonzalez, V. D., Lorang, C. G., Bjornson, Z., Spitzer, M. H., Bodenmiller, B., Fantl, W. J., Pe'Er, D., and Nolan, G. P. (2015). Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nature Protocols*, 10(2):316–333.

Computational Modeling and Analysis of Single-cell Signaling Dynamics in Heterogeneous Cell Populations

James D. Wade

158 Pages

Directed by Dr. Eberhard O. Voit

Cell signaling pathways are complex biochemical systems at the core of cellular information processing. The dynamics of these signaling systems in response to internal and extracellular cues plays a critical role for proper cell functioning. While we have learned much about signaling at the cell population level, no two cells are the same, and cell-to-cell variability can have complex and important consequences for signaling in both individual cells and the cell population as a whole. In many contexts, cells perform essentially identical functions despite their differences, whereas in other contexts, especially in cancer, cell-cell differences in state propagate to differences in function.

The overall goal of this dissertation was the creation of mathematical and computational tools for the study of cell-to-cell variation in signaling and to use these tools to increase our understanding of when single cell differences do, or do not, make a meaningful difference. To address this goal we designed new methods of single-cell analysis, including a computational framework termed single-cell ordinary differential equation modeling (SCODEM) that overcomes the prior experimental trade-off between continuous and multiplexed single-cell measurements of signaling. We tested SCODEM against increasingly demanding datasets, which were all represented in a satisfactory fashion. After the initial analysis of cell-to-cell variability, we analyzed targeted inhibition, protein overexpression and an epithelial-mesenchymal transition. Throughout this process, we provided illustrative examples of how our modeling framework may be used to identify operating principles and limits of signaling systems, which is a first step toward proposing novel therapeutic targets.

The work presented here provides new tools for analyzing cellular heterogeneity and increases our understanding of how differences in cell state effect function by showing intracellular signaling is primarily deterministic at the single cell level. The application of these tools to the dramatic phenotype shift during an epithelial-mesenchymal transition in murine breast cancer cells confirmed that stochasticity plays a much smaller role than had been assumed and that cells modulate signaling without the need of rewiring their signaling network.